

Evaluation of Gaussian approximations for data assimilation in reservoir models

Marco A. Iglesias · Kody J.H. Law · Andrew M. Stuart.

Abstract The Bayesian framework is the standard approach for data assimilation in reservoir modeling. This framework consists mainly in characterizing the posterior distribution of geologic parameters given a prior distribution and data from the reservoir dynamics. Since the posterior distribution quantifies the uncertainty in the geologic parameters of the reservoir, the characterization of the posterior is fundamental for the optimal management of reservoirs. Unfortunately, due to the large-scale highly-nonlinear properties of standard reservoir models, characterizing the posterior is computationally prohibitive. Instead, more affordable *ad hoc* techniques, based on Gaussian approximations, are often used for characterizing the posterior distribution. Evaluating the performance of those Gaussian approximations is typically conducted by assessing their ability at reproducing the truth within the confidence interval provided by the *ad hoc* technique under consideration. This has the disadvantage of mixing-up the approximation properties of the history matching algorithm employed with the information content of the particular observations used, making it hard to evaluate the effect of the *ad hoc* approximations alone.

In this paper we propose to numerically assess the performance of standard Gaussian approximations to probe the Bayesian posterior distribution. In particular we assess the performance of (i) the linearization around the maximum a posterior estimate, (ii) the randomized maximum likelihood and (iii) standard ensemble Kalman filter-type methods. In order to fully resolve the posterior distribution we implement

A.M. Stuart
University of Warwick
Tel.: +44 (0)24 7652 2685
E-mail: A.M.Stuart@warwick.ac.uk

M. A. Iglesias
University of Warwick
Tel.: +44 (0)24 7657 4827
E-mail: M.A.Iglesias-Hernandez@warwick.ac.uk

K.J.H. Law
University of Warwick
Tel.: +44 (0)24 7652 8332
E-mail: K.J.H.Law@warwick.ac.uk

a state-of-the art MCMC method that scales well with respect to the dimension of the parameter space. Our implementation of the MCMC method provides the gold standard against which to assess the aforementioned Gaussian approximations. We present numerical synthetic experiments where we quantify the capability of each of the *ad hoc* Gaussian approximation in reproducing the mean and the variance of the posterior distribution (characterized via MCMC) associated to a data assimilation problem. Both single-phase and two-phase (oil-water) reservoir models are considered so that fundamental differences in the resulting forward operators are highlighted. The main objective of our controlled experiments is to exhibit the substantial discrepancies of the approximation properties of standard *ad hoc* Gaussian approximations. Numerical investigations of the type we present here will lead to greater understanding of the cost-efficient, but *ad hoc*, Bayesian techniques used for data assimilation in petroleum reservoirs, and hence ultimately to improved techniques with more accurate uncertainty quantification.

Keywords First keyword · Second keyword · More

1 Introduction

Simulating the dynamics of a reservoir involves solving a large-scale numerical model that depends on parameters related to petrophysical properties of the reservoir. These properties need to be known at each discretization point of the physical domain of the reservoir. Unfortunately, direct measurements of petrophysical properties are only available at a small number of locations within the domain of interest. Therefore, a statistical description of the reservoir parameters is required to properly account for the uncertainty in the petrophysical properties caused by the lack of information. A *prior distribution* of geologically consistent reservoir parameters can be generated, for example, from a variogram analysis conducted on static data from core samples. Geostatistical techniques can then be used to generate realizations of reservoir parameters conditioned to static data [8]. In some cases, information concerning geologic facies may also be incorporated [2]. On the other hand, with the aid of downhole permanent sensors, measurements of reservoir flow can be continuously acquired. In a Bayesian framework, these flow measurements, the reservoir model and the prior distribution of the petrophysical properties are combined to characterize the *posterior distribution* of the reservoir parameters given dynamic (flow) data. This posterior distribution quantifies the uncertainty in the reservoir predictions and it is essential for assessing the economical and environmental risk of oil recovery procedures.

Markov Chain Monte Carlo (MCMC) methods are the standard techniques for sampling the posterior distribution described above. In particular, the Metropolis-Hasting variant of MCMC has been typically used for data assimilation in reservoir models [3, 10, 19, 22, 17, 9, 12]. In general, the posterior distribution that arises from Bayesian data assimilation does not admit a finite-dimensional parametrization, with the exception of very few particular cases such as the linear and Gaussian case. Therefore, strictly speaking, an infinite number of samples are required to define it. This implies, in practice, that hundreds of thousands or even millions of reservoir simulations may be required for standard MCMC methods to accurately characterize

the posterior distribution. This computational disadvantage of standard MCMC approaches has given rise to the development of more computationally efficient MCMC techniques [7, 9, 19, 12]. With increasing advance of computational power, the aforementioned MCMC methods may potentially become viable tools for reservoir management in the decades to come. At the present time, however, it is essential to develop techniques that provide a reasonable characterization of the posterior with a computational cost that involves a limited number of reservoir model runs. It then comes as no surprise that, in the previous years, research on data assimilation for uncertainty quantification (UQ) applications has mainly focused on improving the efficiency and accuracy of *ad-hoc* ensemble-based techniques that provide approximations of the posterior distribution based on Gaussian assumptions.

As pointed out in the recent literature review of [21], there are three main approaches that have been consistently adopted for sampling approximations of the posterior distribution: (i) linearization around the maximum a posteriori (MAP) estimate (LMAP), (ii) randomized maximum likelihood (RML) method and (iii) ensemble Kalman filter (EnKF). Under a Gaussian prior and a linear model, it can be shown that all these techniques provide samples of the posterior distribution [20, 16]. We therefore refer to those methods as *Gaussian approximations* of the posterior. For standard (nonlinear) reservoir models, the mathematical structure of the approximation provided by the three approaches is still unknown. Nonetheless, the aforementioned methods are widely applied for generating model parameters conditioned to dynamics data, which are then used for statistical analysis of reservoir performance. Consequently, in the Bayesian framework, optimal decision-making and risk management depend on the quality of the underlying Gaussian approximations of the posterior. It is therefore, fundamental to understand the accuracy and convergence properties of those Gaussian approximations in order to interpret predictions of uncertainty made using them, and in order to develop improved methodologies from them. Rigorous numerical studies of these Gaussian approximate algorithms can shed light on these issues and can point us towards theoretical analyses. In this paper we therefore provide a numerical evaluation of the performance of the *ad hoc* Gaussian approximate algorithms LMAP, RML, and variants of the EnKF methodology, by using an expensive, but full resolved, MCMC simulation as our gold standard. This approach is analogous to the recent study of similar Gaussian approximate algorithms arising in the context of atmospheric data assimilation [15].

1.1 Literature review

Although the theoretical aspects of the approximation properties of LMAP, RML and EnKF are unknown for the case of nonlinear models, some numerical investigations have been performed [17, 3]. To our best knowledge, only the work in [17] provides an evaluation of approximate methods for sampling the posterior. To accomplish this goal, Liu et-al [17] use a standard random walk MCMC method to generate accurate samples of the posterior. These, in turn, are used as gold-standard against to which compare the performance of the approximate methods: LMAP, RML and pilot point methods. In their evaluation, Liu et-al use synthetic data from a single-phase one-

dimensional reservoir discretized with 20 gridblocks. The main conclusion of [17] is that RML provides the best uncertainty quantification when compared against the MCMC gold standard. In particular, RML outperforms pilot point methods whose application to reservoir data assimilation problems has been lately abandoned.

Within the context of evaluating the uncertainty quantification properties of data assimilation techniques, it is relevant to mention the work of [3] where several methods were compared for the synthetic PUNQ-S3 reservoir model. The main goal of [3], however, is to evaluate the ability of the corresponding techniques to provide confidence of interval that contain the truth estimate. Among those techniques, MCMC is the only one that can potentially provide accurate samples from the Bayesian posterior. Unfortunately, as stated in [3], the MCMC results are not conclusive due to the small chain used for their experiments. Thus, the evaluation in [3] does not provide an evaluation in the strict sense of a Bayesian framework. It is worth mentioning that the work of [3] and [17] appeared almost a decade ago when EnKF had just been introduced to the history matching community [1], and so EnKF was not assessed in [17, 3]. Almost a decade later and after hundreds of publications, EnKF is now perhaps the only computationally feasible technique for real-time data assimilation in petroleum reservoirs. For a comprehensive review of EnKF for reservoir applications, we refer the reader to [1].

In the recent work of [12], the EnKF is combined with an MCMC algorithm to improve the efficiency of standard MCMC methods for sampling the posterior in a Bayesian framework. Although the analysis of the approximate properties of EnKF is not the main goal of [12], an implicit evaluation of EnKF is displayed. Indeed, under the assumption that the MCMC samples of [12] provide an accurate characterization of the posterior, then [12] provides a partial assessment of EnKF for approximating the posterior. While the aforementioned work exposes severe limitations of the EnKF for sampling the posterior, an evaluation with respect to other approximate methods remains nonexistent.

In the context of evaluating the uncertainty quantification properties of Gaussian approximations of the posterior, we highlight the work of [13, 25]. For the PUNQ-S3 model mentioned above, [13] compares the performance of EnKF and RML. In [25], a new SVD-based RML is introduced and compared against EnKF. It is important to mention, however, that [13, 25] evaluate the capability of these Gaussian approximations for reproducing the truth within the confidence of interval of the technique under consideration. While this is a natural strategy for assessing uncertainty quantification properties, it is an insufficient evaluation from the perspective of the Bayesian framework. In other words, capturing the truth within the spread of model predictions obtained with a Gaussian approximation does not ensure that the spread correctly represents the uncertainty quantified by the posterior distribution of Bayesian data assimilation. It is therefore essential to develop a controlled experiment where standard Gaussian approximation can be tested against the solution to the Bayesian data assimilation problem: the posterior distribution.

1.2 The proposed work

In this paper we propose the numerical evaluation of LMAP, RML and some standard versions of ensemble Kalman filter-type methods for approximating the posterior distribution within the Bayesian framework of data assimilation. We characterize the posterior distribution by using a state-of-the art MCMC method that provides a gold-standard against which to compare the aforementioned Gaussian approximations. In this sense, our work has a similar goal to the one of [17]. However, there are two recent algorithmic developments which motivate our desire to revisit the perspective introduced in [17]. The first, discussed in detail below, is that MCMC methodology has evolved significantly, enabling the study of considerably more sophisticated forward models and more high dimensional parameterizations of the unknown petrophysical quantities, leading to greater realism. The second new aspect of our work is the assessment of ensemble Kalman filter-type methods. In particular we consider the most standard EnKF implementations, namely: (i) the perturbed observation EnKF; and (ii) the square root filter EnSRF of [23, 10]. In both cases we also evaluate the effect of performing distance-based localization [5, 11].

We emphasize that, in contrast to other approaches [13, 25, 3, 5, 11] where the aim is to recover the truth within the confidence interval of relevant quantities and to history-match data, here we are interested in assessing the performance for characterizing the posterior distribution. Evaluation of algorithms by their ability to recover the truth within a confidence interval has the disadvantage of entangling the approximation properties of the history matching algorithm employed with the information content of the particular observations used, making it hard to evaluate the effect of the *ad hoc* approximations alone. Our assessment of the ability to probe the Bayesian posterior distribution is conducted by quantifying the capability of each Gaussian approximation in reproducing the mean and the variance of the posterior distribution associated to a data assimilation problem. Two prototypical reservoir models are used, both in two spatial dimensions: (i) slightly compressible single-phase Darcy flow model and (ii) incompressible oil-water reservoir model. In both models, the unknown is the logarithm of the absolute permeability of the reservoir $u = \log K$. For the single-phase model, pressure data is collected from production wells. For the oil-water model, total flow rate is measured at the production wells while bottom-hole pressure is collected at the injection wells. For both models considered here, the corresponding parameter-to-output map $G(u)$ is nonlinear. Thus, even when the prior distribution is Gaussian, the Bayesian posterior is non-Gaussian. This constitutes the ideal scenario to evaluate approximation properties of the techniques of interest, provided that a gold standard is obtained from accurately sampling the posterior as we describe below.

As we indicated earlier, some MCMC methods have been used for sampling the Bayesian posterior distribution in reservoir models. However, some of these methods [9, 19] rely on reducing the parameter space (e.g. via truncating Karhunen-Loeve expansion) and/or upscaling the model to reduce the computational cost of the algorithm. In the present experiment, however, we are interested in the more general case where no reduction of the parameter space is possible. In other words, we assume that the petrophysical property is unknown at each at the location of the phys-

ical domain of the reservoir. In this case, a standard MCMC technique like the one used in [17] is computationally prohibitive for larger size problems like the ones considered here. To overcome this difficulty, we take advantage of recent developments in MCMC methodology and sample the posterior by applying the preconditioned Crank-Nicolson (pCN) MCMC method described in [7], and derived from the infinite-dimensional Bayesian framework developed in [24]. In contrast to standard MCMC methods, the acceptance probability in the pCN-MCMC method is invariant with respect to the dimension of the parameter space, therefore making pCN-MCMC ideal for large-scale problems like the one studied here. The advantage of using pCN-MCMC over standard MCMC for data assimilation in some geophysical problems has been shown in [7]. For petroleum reservoir applications, the computational efficiency of pCN-MCMC with respect to other existing methods deserves further investigation. Nevertheless, in the present work we apply pCN-MCMC and provide numerical evidence of convergence so that the corresponding realizations generated with pCN-MCMC are samples from the Bayesian posterior. These, in turn, provide a gold-standard against to which compare LMAP, RML and standard versions of ensemble Kalman filter-type of methods. The proposed numerical evaluation of the approximation techniques has two concrete goals: (i) assess the capability to recover the mean and variance of the posterior and (ii) evaluate the performance for reproducing the uncertainty (quantified by the posterior) in the reservoir model predictions.

In Section 2 we describe the prototypical reservoir models that define the forward operators that we use for data assimilation. The Bayesian framework for data assimilation as well as the MCMC methodology for sampling the posterior are introduced in Section 3. Methodologies based on Gaussian approximations of the posterior are described in Section 4. In Section 5 we report and discuss the numerical results and comparisons of our synthetic experiments. The summary and final remarks are presented in Section 6.

2 Forward Reservoir Models

In this section we briefly outline the forward (reservoir) models that we use for the evaluation of Gaussian approximations of the posterior. On the one hand, we consider simplified two-dimensional models for which a forward model run is computationally inexpensive and therefore feasible for the highly computationally challenging MCMC method. On the other hand, by sharing the mathematical structure of more sophisticated models, the models we describe below are ideal for prototyping and evaluating performance in a controlled fashion. For each of the following models, we consider a two-dimensional reservoir whose physical domain, absolute permeability and porosity are denoted by D , K and ϕ respectively. The interval $[0, T]$ ($T > 0$) is the time interval of interest for the flow simulation. For each reservoir model, we define the forward operator $G : X \rightarrow \mathbb{R}^N$ that maps the parameter space X into the observation space \mathbb{R}^N . In other words, $G(u)$ is the model predictions corresponding to the parameter $u \in X$. For simplicity we assume that the only unknown parameter is $u = \log K$. Nevertheless, all the techniques and implementations that we describe in

subsequent sections can be extended to include additional parameters, such as porosity which is routinely estimated alongside permeability in many practical scenarios.

2.1 Single-phase Darcy flow

We consider a single-phase reservoir where oil is produced at N_w production wells operated under prescribed production rates $\{q^l(t)\}_{l=1}^{N_w}$ ($t \in [0, T]$). The flow in the reservoir is described in terms of the (state variable) fluid pressure $p(x, t)$ ($(x, t) \in D \times [0, T]$) which is governed by the following equation [6]

$$c\phi \frac{\partial p}{\partial t} - \nabla \cdot e^u \nabla p = \sum_{l=1}^{N_w} q^l \delta(x - x^l) \quad \text{in } D \times (0, T] \quad (1)$$

where $u \equiv \log K$, c is the total compressibility and $\delta(x - x^l)$ is (a possibly mollified) Dirac delta centered at the l -th well with location denoted by x^l . In addition, we consider the following boundary and initial conditions

$$-e^u \nabla p \cdot \mathbf{n} = 0 \quad \text{on } \partial D \times (0, T], \quad (2)$$

$$p = p_0 \quad \text{in } D \times \{0\}. \quad (3)$$

As we indicated earlier, the only uncertain parameter in (1)-(3) is the log-permeability u . Therefore, the additional model parameters c , ϕ , v , p_0 and the geometry D in (1)-(3) are prescribed.

In order to construct the forward operator, we first define the model predictions of measurements. Let us then assume that N_M measurements of pressure from wells are collected at times t_1, \dots, t_M . We define the measurement functional

$$M_n^l(p) = p(x^l, t_n) \quad (4)$$

that corresponds to the fluid pressure at time t_n and well location x^l . For the exposition of subsequent sections we also define the vector

$$M_n(p) = (M_n^1(p), \dots, M_n^{N_w}(p)). \quad (5)$$

We finally define $N = N_w N_M$, i.e. the total number of observations from wells, and construct the forward operator

$$G(u) = (M_1(p), \dots, M_{N_M}(p)) \quad (6)$$

Note that p in (6) depends on u via (1)-(3).

2.2 Oil-water reservoir model

We consider an oil-water reservoir model initially saturated with oil and irreducible water. Let us index by $\gamma = w$ and $\gamma = o$ the water and oil phase respectively. We assume that both fluids and the rock are incompressible. We are interested in a waterflood process where water is injected at N_I injection wells located at $\{x_I^l\}_{l=1}^{N_I}$. Water and oil are produced at N_P production wells located at $\{x_P^l\}_{l=1}^{N_P}$. Additionally, we assume that injection wells are operated under prescribed rates $\{q^l(t)\}_{l=1}^{N_I}$ while production wells are constrained to bottom hole pressure denoted by $\{P_{bh}^l(t)\}_{l=1}^{N_P}$. The reservoir dynamics in $[0, T]$ are described by the (state variables) water saturation and the pressure denoted by $s(x, t)$ and $p(x, t)$ respectively $((x, t) \in D \times [0, T])$. From standard arguments it can be shown that (s, p) is the solution to the following system [6]

$$-\nabla \cdot \lambda(s) e^u \nabla p = \sum_{l=1}^{N_I} q^l \delta(x - x_I^l) + \sum_{l=1}^{N_P} \omega^l \lambda(s) [P_{bh}^l - p] \delta(x - x_P^l), \quad (7)$$

$$\phi \frac{\partial s}{\partial t} - \nabla \cdot \lambda_w(s) e^u \nabla p = \sum_{l=1}^{N_I} q^l \delta(x - x_I^l) + \sum_{l=1}^{N_P} \omega^l \lambda_w(s) [P_{bh}^l - p] \delta(x - x_P^l), \quad (8)$$

in $D \times (0, T]$, where $\delta(x - x_I^l)$ and $\delta(x - x_P^l)$ are the (possibly mollified) Dirac deltas as defined before, $\{\omega^l\}_{l=1}^{N_P}$ are constants related to the well model [6]. Additionally $\lambda_w(s)$ and $\lambda(s)$ denote the water and total mobility defined by

$$\lambda_w(s) = \frac{k_{rw}(s)}{\mu_w}, \quad \lambda(s) = \frac{k_{ro}(s)}{\mu_o} + \lambda_w(s) \quad (9)$$

where $k_{r\gamma}(s)$ and μ_γ denote the relative permeability and the viscosity of the γ -phase fluid, respectively. Furthermore, we assume that

$$k_{rw}(s) = a_w \left[\frac{s - s_{iw}}{1 - s_{iw} - s_{or}} \right]^2, \quad k_{ro}(s) = a_o \left[\frac{1 - s - s_{or}}{1 - s_{iw} - s_{or}} \right]^2 \quad (10)$$

where $a_w, a_o \in (0, 1]$, s_{iw} is the irreducible water saturation and s_{or} is the residual oil saturation. We additionally prescribe initial conditions for pressure and water saturation

$$p = p_0, \quad s = s_0 \quad \text{in } D \times \{0\} \quad (11)$$

For simplicity, no-flow boundary conditions are prescribed on the reservoir boundary

$$-e^u \lambda(s) \nabla p \cdot \mathbf{n} = 0 \quad \text{on } \partial D \times (0, T], \quad (12)$$

$$-e^u \lambda_w(s) \nabla p \cdot \mathbf{n} = 0 \quad \text{on } \partial D \times (0, T]. \quad (13)$$

Let us assume that there are N_M measurement times denoted as before $\{t_n\}_{n=1}^{N_M}$. We assume measurements of bottom-hole pressure are collected at the injection wells at $\{t_n\}_{n=1}^{N_M}$. This, according to Peaceman's well-model [6], is defined by

$$M_n^{l,I}(p, s) = P_{bh}^{l,I}(t_n) = \left[\frac{q^l(t_n)}{\omega^l \lambda(s(x_I^l, t_n))} + p(x_I^l, t_n) \right] \quad (14)$$

for $l = 1, \dots, N_I$ and $n = 1, \dots, N_M$. Analogously, we consider measurements of total flow rate at the production wells

$$M_n^{l,P}(p, s) = q^{l,P}(t_n) = \omega^l \lambda(s(x_P^l, t_n)) [P_{bh}^l(t_n) - p(x_P^l, t_n)] \quad (15)$$

for $l = 1, \dots, N_P$ and $n = 1, \dots, N_M$. Let us denote by $N_w = N_P + N_I$ the total number of wells, and define the N_w -dimensional vector

$$M_n(p, s) = (M_n^{1,I}(p, s), \dots, M_n^{N_I,I}(p, s), M_n^{1,P}(p, s), \dots, M_n^{N_P,P}(p, s)) \quad (16)$$

The total number of measurements N is defined as before and the forward map $G : X \rightarrow \mathbb{R}^N$ is then given by expression

$$G(u) = (M_1(p, s), \dots, M_{N_M}(p, s)) \quad (17)$$

which in this case comprises the production data obtained from production and injection wells at the measurement times.

With both forward models written in terms of the forward operator G , in the next section we describe the Bayesian inverse problem of finding u given noisy observations of $G(u)$.

3 The Bayesian framework

We assume that the unknown parameter u and the data $y \in Y$ are related by

$$y = G(u) + \eta \quad (18)$$

where G is the forward map introduced in the previous section, and $\eta \in \mathbb{R}^N$ is a vector of random noise. Informally the Bayesian approach to inversion proceeds by placing a *prior* probability distribution $\mathbb{P}(u)$ on u and assuming an independent probability distribution on η . The *likelihood*, namely the probability of the observed data y given a particular instance of the unknown parameter u , is then denoted $\mathbb{P}(y|u)$. Bayes' rule then states that the *posterior* probability of the unknown parameter u , given the observed data y , denoted by $\mathbb{P}(u|y)$, is determined by the formula

$$\frac{\mathbb{P}(u|y)}{\mathbb{P}(u)} \propto \mathbb{P}(y|u). \quad (19)$$

In this section we formulate this precisely in the case where the unknown parameter is a function.

We define the norm $\|\cdot\|_B = \|B^{-1/2}(\cdot)\|$ for any covariance operator B and we use this notation throughout the paper, in particular in the observation space, with $B = \Gamma$, and in the log-permeability space with $B = C$. For simplicity and following convention in the field, we will not distinguish notationally between the random variable and its realization, except in the case of the truth, which will be important to distinguish by u^\dagger in subsequent sections in which it will be prescribed and known.

3.1 An infinite-dimensional Bayesian framework

We are interested in the inverse problem of characterizing the posterior distribution of the unknown log-permeability *function* u given *finite-dimensional* observational data denoted by y . We approach this inverse problem by means of the infinite-dimensional Bayesian framework of [24] that we now briefly describe. Assume that $u \in X$ where X is a Hilbert space, and denote by μ_0 the prior probability measure on u . We assume that the unknown u and the data $y \in Y$ are related by (18). For simplicity, we assume that $\eta \sim N(0, \Gamma)$. Then, the rigorous interpretation of (19) is that the posterior distribution on $u|y$ is given by measure μ satisfying

$$\frac{d\mu}{d\mu_0}(u) = \frac{\exp(-\Phi(u, y))}{\int_X \exp(-\Phi(u, y)) \mu_0(du)} \quad (20)$$

where the left hand side of (20) is the Radon-Nikodym derivative of the posterior distribution $\mu(u) = \mathbb{P}(u|y)$ with respect to the prior μ_0 and

$$\Phi(u, y) = \frac{1}{2} \|y - G(u)\|_F^2.$$

A sufficient condition for this to be well-defined is that $\Phi(\cdot, y)$ is continuous as a mapping from X into \mathbb{R} for each fixed y , and that $\mu_0(X) = 1$ so that functions drawn from μ_0 are in X almost surely. The formula (20) then holds in infinite-dimensions, exhibiting the posterior density with respect to the prior; in practical terms this means that posterior expectations can be found by reweighting prior expectations by the right-hand side of (20).

The posterior distribution μ quantifies the uncertainty of the logarithm of the absolute permeability given production data, normalized by the prior. Since G is non-linear, the posterior is non-Gaussian even when the prior μ_0 is Gaussian. Thus there is no useful closed-form expression for the posterior distribution and it must be characterized by means of sampling.

Before we describe the approach for sampling μ , we note that, if we assume the prior μ_0 is Gaussian with mean \bar{u} and covariance C , it follows that the maximum a posteriori (MAP) estimate u_{MAP} is the minimizer of the functional

$$J(u) = \Phi(u, y) + \frac{1}{2} \|u - \bar{u}\|_C^2 \quad (21)$$

The MAP estimator is the typical estimate computed in standard history matching problems where the goal is to recover the truth by fitting historic production data. Note that the Bayesian approach thus subsumes this classical approach to inversion, whilst also providing rigorous quantification of uncertainty of predictions, given clear assumptions on the prior and noise probabilities.

3.2 Sampling the posterior with MCMC

A state-of-the-art class of MCMC methods that sample from μ defined in (20) has been proposed in [7]. In particular, we consider the following preconditioned Crank-Nicolson (pCN) MCMC [7, Section 5.2].

Algorithm 1 (pcN-MCMC) Take $u^{(0)} \sim N(\bar{u}, C)$, $n = 1$, and $\beta \in (0, 1)$. Then,

(1) *pcN proposal. Generate u from*

$$u = \sqrt{1 - \beta^2} u^{(n)} + (1 - \sqrt{1 - \beta^2}) \bar{u} + \beta \xi, \quad \text{with } \xi \sim N(0, C) \quad (22)$$

(2) *Set $u^{n+1} = u$ with probability $a(u^n, u)$ and $u^{n+1} = u^n$ with probability $1 - a(u^n, u)$, where*

$$a(u, v) = \min \left\{ 1, \exp(\Phi(u, y) - \Phi(v, y)) \right\} \quad (23)$$

(3) *$n \mapsto n + 1$ and repeat.*

All the probabilities are generated independently of one another, leading to a Markov chain which is invariant with respect to μ . Notice that the small change in proposal, when compared with the standard random walk MCMC [17], results in an acceptance probability defined via differences of Φ and not J . Because Φ is finite with respect to μ , whilst J is not, this leads to a considerably improved algorithm which has desirable $\dim(X)$ -independent properties when implemented on a sequence of approximating problems with $\dim(X) \rightarrow \infty$. Therefore, for large $\dim(X)$ like the one considered here, pcN-MCMC provides a more robust and efficient technique than the standard MCMC approaches. Numerical evidence of these scaling properties can be found in [7].

Based on the forward models of Section 2, the purpose of the present work is to design synthetic experiments for solving the Bayesian data assimilation problem, i.e. finding the posterior distribution. By implementing the pcN-MCMC algorithm, we characterize this posterior and generate a gold standard against to which compare the Gaussian approximations that we introduce in the following section.

4 Gaussian approximations of the posterior

In this section we introduce some standard *ad-hoc* methods that use Gaussian approximations to sample the posterior distribution (20). In particular, we consider LMAP, RML, EnKF and EnSRF which have been typically used for history matching and uncertainty quantification in the Bayesian framework of data assimilation of petroleum reservoirs. While many variants of the aforementioned techniques can be found in the literature [21], here we focus on the most standard and typical implementations used for history matching. For each of the aforementioned techniques, the objective of the subsequent description is twofold. First, we introduce the algorithm and the associated computational cost. Second, we indicate the type of Gaussian approximation made for the definition of the technique under consideration.

4.1 Linearization around the MAP (LMAP)

As described in Section 3, the minimizer of J introduced in (21) defines the MAP estimator, i.e.

$$u_{MAP} = \operatorname{argmin}_u \left\{ \Phi(u, y) + \frac{1}{2} \|u - \bar{u}\|_C^2 \right\} \quad (24)$$

We can further define,

$$C_{MAP} = C - CQ^T(QCQ^T + \Gamma)^{-1}QC \quad (25)$$

where $Q \equiv DG(u_{MAP})$ is the Frechet derivative of G evaluated at $u = u_{MAP}$. The linearization around the MAP [20, Section 10.5] consists of approximating the posterior μ in (20) by $\mu \approx N(u_{MAP}, C_{MAP})$.

The LMAP algorithm approximates the posterior with an ensemble of N_{en} realizations from $N(u_{MAP}, C_{MAP})$ [20, Section 10.5]. This ensemble can then be used to approximate integrals with respect to the posterior of *nonlinear* functions of u . Note that when G is linear, $\mu = N(u_{MAP}, C_{MAP})$ and then u_{MAP} and C_{MAP} are the mean and covariance of the posterior. The algorithm, however, is well-defined in general, and may thus be applied to cases in which G is nonlinear.

Algorithm 2 (LMAP) (1) Compute u_{MAP} and C_{MAP} from (24) and (25) respectively.
 (2) Compute the Cholesky factor L of C_{MAP} , i.e. $C_{MAP} = LL^T$.
 (3) For $j \in \{1, \dots, N_{en}\}$, generate

$$u^{(j)} = u_{MAP} + L^T z^{(j)} \quad (26)$$

where $z^{(j)} \sim N(0, I)$.

Samples generated by (26) are draws from $N(u_{MAP}, C_{MAP})$ and so the ensemble $\{u^{(j)}\}_{j=1}^{N_{en}}$ provides an approximation to $N(u_{MAP}, C_{MAP})$ and, hence the posterior.

The computational cost of LMAP depends on the cost of computing the MAP estimator (24) and the factorization of C_{MAP} . For the present work, we develop implementations of the Levenberg-Marquardt algorithm of [20, Section 8.4] with the stopping criteria given by (8.82) and (8.83) from [20, Section 8.5]. It is worth mentioning that, within the context of reservoir characterizations, multiple techniques for computing the MAP estimator have been widely studied (e.g. BFGS, LBFGS, Gauss-Newton) [20, Section 8]. It is of interest to evaluate the optimal minimization technique, but this is beyond the scope of our present work.

4.2 Randomized Maximum Likelihood (RML)

The RML technique was developed as an attempt to accelerate MCMC methods for sampling the posterior from Bayesian data assimilation in reservoir models [22]. The main idea of RML is to construct an ensemble of MAP estimators from randomized objective functions (24). A standard implementation of RML is presented in the following algorithm

Algorithm 3 (RML) For $j \in \{1, \dots, N_e\}$

- (1) Generate $u^{(j)} \sim N(\bar{u}, C)$
- (2) Define $y^{(j)} = y + \eta^{(j)}$ with $\eta^{(j)} \sim N(0, \Gamma)$.
- (3) Compute

$$u_{RML}^{(j)} = \operatorname{argmin}_u \left\{ \Phi(u, y^{(j)}) + \frac{1}{2} \|u - u^{(j)}\|_C^2 \right\}. \quad (27)$$

In the case where G is linear, the RML algorithm can be shown to sample the posterior distribution (i.e. from $\mu = N(u_{MAP}, C_{MAP})$) [16]. For the nonlinear case of interest here, the RML algorithm provides an approximation the nature of which, to the best of our knowledge, has not been systematically understood.

Note that for each ensemble member, RML requires the solution to the minimization problem (27). Nevertheless, since each minimization problem is independent from one another, RML is then embarrassingly parallelizable. Each of those minimization problems has the same structure as the one that we solve for the MAP estimator (24). For a relatively small problem the computational cost of computing L in LMAP, given Q which has already been constructed while solving (24), as well as the generation of (26), are negligible compared to the cost of one forward model evaluation. Thus the computational cost of RML is roughly N_e times the computational cost of LMAP, although the effect of the multiplier N_e can be ameliorated in a parallel context.

Similarly to our implementation of the MAP, for RML we consider the Levenberg-Marquardt method and the corresponding stopping criteria mentioned above. Improving the optimization technique required for (27) can reduce the overall computational cost of RML. Alternative methods to reduce the computational cost of RML by means of a truncated SVD approach can be found in [25].

4.3 Ensemble Kalman Filter (EnKF)

Ensemble methods based on the Kalman filter have been extensively applied for Bayesian data assimilation in petroleum reservoir applications. For a complete review of most of the EnKF implementations we refer the reader to the monograph of [1]. In this section we briefly discuss some relevant aspects of EnKF in the context of history matching of petroleum reservoirs. These ensemble Kalman filter-type of algorithms, make Gaussian approximations in a sequential manner as we describe below. As a result, for the general case, those techniques do not provide correct sampling of the posterior (20). Nevertheless, due to its ease of implementation and low computational cost, ensemble Kalman filter-type of methods are arguably the only feasible techniques for online data assimilation in subsurface applications.

4.3.1 Introduction and Main Algorithm

In order to introduce the algorithms, we first consider a sequential formulation of the reservoir model. In particular, let us define v_n the state variable at time t_n and S the

state space. For example, for the single-phase model of Section 2, $v_n = p(x, t_n)$. We define the solution operator $\Psi_n : S \times X \rightarrow S$

$$v_n = \Psi_n(v_{n-1}, u) \quad (28)$$

which, for a given parameter u , maps the state variable from time $t = t_{n-1}$ to $t = t_n$. In practice, Ψ_n is simply the numerical solver that arises from the time discretization of the reservoir model under consideration. In addition, we assume that data is given at each of these points in time and is correlated between times only through the state itself, i.e.

$$y_n = M_n(v_n) + \eta_n, \quad (29)$$

where $\eta_n \sim N(0, \Gamma_n)$ and $M_n : S \rightarrow \mathbb{R}^{N_w}$ is the measurement functional acting on the state variable at time $t = t_n$. For the models of Section 2, M_n is defined by (5) and (16) respectively and N_w is the number of total wells. Define,

$$z = \begin{pmatrix} u \\ v \\ w \end{pmatrix}, \quad \Xi_n(z) = \begin{pmatrix} u \\ \Psi_n(v, u) \\ M_n(\Psi_n(v, u)) \end{pmatrix}. \quad (30)$$

Since the permeability in the forward reservoir model does not change in time, it follows that (28)-(29) can be written as

$$z_n = \Xi_n(z_{n-1}), \quad (31)$$

$$y_n = H z_n + \eta_n. \quad (32)$$

where $H = (0, 0, I)$. We now consider the following standard perturbed observation version of EnKF [1].

Algorithm 4 (EnKF) *Construct an initial ensemble*

$$z_0^{(j,a)} = \begin{pmatrix} u_0^{(j)} \\ v_0 \\ M_0(v_0) \end{pmatrix} \quad (33)$$

where $\{u_0^{(j)}\}_{j=1}^{N_e} \sim \mu_0$ and v_0 is the initial condition for the state variable. For $j = 1, \dots, N_m$

(1) *Prediction Step: Propagate the ensemble of particles forward under (31) giving*

$$z_n^{(j,f)} = \Xi_n(z_{n-1}^{(j,a)}) \quad j \in \{1, \dots, N_e\} \quad (34)$$

From this ensemble we define a sample mean and covariance as follows:

$$\bar{z}_n^f = \frac{1}{N_e} \sum_{j=1}^{N_e} z_n^{(j,f)} \quad (35)$$

$$C_n^f = \frac{1}{(N_e-1)} \sum_{j=1}^{N_e} z_n^{(j,f)} (z_n^{(j,f)})^T - \bar{z}_n^f (\bar{z}_n^f)^T \quad (36)$$

(2) *Analysis step: Compute the updated ensembles*

$$z_n^{(j,a)} = z_n^{(j,f)} + K_n(y_n^{(j)} - H z_n^{(j,f)}) \quad (37)$$

where

$$K_n = C_n^f H^T (H C_n^f H^T + \Gamma_n)^{-1} \quad (38)$$

and

$$y_n^{(j)} = y_n + \eta_n^{(j)}, \quad \eta_n^{(j)} \sim N(0, \Gamma_n) \quad (39)$$

Here the $\eta_n^{(j)}$ are chosen i.i.d. In order to discuss the computational cost of the EnKF algorithm, let us first note that all the vectors and matrices involved have block structure inherited from the structure of the space $Z = X \times S \times \mathbb{R}^{N_w}$. For example, we have

$$z_n^{(j,f)} = \begin{pmatrix} u_n^{(j,f)} \\ v_n^{(j,f)} \\ w_n^{(j,f)} \end{pmatrix} = \begin{pmatrix} u_{n-1}^{(j,a)} \\ \Psi_n(v_{n-1}^{(j,a)}, u_{n-1}^{(j,a)}) \\ M_n(\Psi_n(v_{n-1}^{(j,a)}, u_{n-1}^{(j,a)})) \end{pmatrix}, \quad \bar{z}_n^f = \begin{pmatrix} \bar{u}_n^f \\ \bar{v}_n^f \\ \bar{w}_n^f \end{pmatrix}$$

We also have

$$C_n^{zw,f} = \begin{pmatrix} C_n^{uw,f} \\ C_n^{vw,f} \\ C_n^{ww,f} \end{pmatrix}, \quad C_n^f = \begin{pmatrix} C_n^{uu,f} & C_n^{uv,f} & C_n^{uw,f} \\ (C_n^{uv,f})^T & C_n^{vv,f} & C_n^{vw,f} \\ (C_n^{uw,f})^T & (C_n^{vw,f})^T & C_n^{ww,f} \end{pmatrix}.$$

Then, expression (37) can be written as

$$z_n^{(j,a)} = \Xi_n(z_{n-1}^{(j,a)}) + C_n^{zw,f} (C_n^{ww,f} + \Gamma)^{-1} (y_n^{(j)} - M_n(\Psi_n(v_{n-1}^{(j,a)}, u_{n-1}^{(j,a)}))) \quad (40)$$

The submatrices in C_n^f needed for (40) are given by

$$C_n^{uw,f} = \frac{1}{N_e} \sum_{j=1}^{N_e} u_n^{(j,f)} (w_n^{(j,f)})^T - \bar{u}_n^f (\bar{w}_n^f)^T, \quad (41)$$

$$C_n^{vw,f} = \frac{1}{N_e} \sum_{j=1}^{N_e} v_n^{(j,f)} (w_n^{(j,f)})^T - \bar{v}_n^f (\bar{w}_n^f)^T, \quad (42)$$

$$C_n^{ww,f} = \frac{1}{N_e} \sum_{j=1}^{N_e} w_n^{(j,f)} (w_n^{(j,f)})^T - \bar{w}_n^f (\bar{w}_n^f)^T, \quad (43)$$

We recall that $w \in \mathbb{R}^{N_w}$ where N_w is the number of wells. Typically N_w is much smaller than the dimensions of the (discretized) parameter space X . Consequently, the computational cost of constructing C_n^{zw} and C_n^{ww} and inverting the $(C_n^{ww} + \Gamma_n)^{-1}$ in (40) is negligible compared to the cost of computing $\Xi_n(z_{n-1}^{(j,a)})$, which from (30) we can see is mainly determined by the cost of $\Psi_n(v_{n-1}^{(j,a)}, u_{n-1}^{(j,a)})$ (i.e. running the reservoir simulator in the time-interval $[t_{n-1}, t_n]$). Therefore, the computational cost of the EnKF is approximately N_e times the cost of a forward model simulation.

4.3.2 Derivation by Gaussian Approximation of the Filtering Distribution

We now indicate how a Gaussian approximation gives rise to the EnKF algorithm presented above. We start by defining the conditional measures for $n_1, n_2 \leq N_m$

$$\mu_{n_1|n_2}(z_{n_1}) = \mathbb{P}(z_{n_1} | \{y_k\}_{k=1}^{n_2}) \quad (44)$$

In the filtering approach, given the prior distribution $\mu_{n|n-1}$ of z_n given data up to the previous time $t = t_{n-1}$ is combined with data provided at the current time $t = t_n$ to define the posterior distribution ($\mu_{n|n}$) of z_n given data up to the current time $t = t_n$. The latter can be obtained from Bayes rule:

$$\frac{\mu_{n|n}(z)}{\mu_{n|n-1}(z)} \propto \exp\{-\Phi_n(z)\} \quad (45)$$

where

$$\Phi_n(z) = \frac{1}{2} \|y_n - Hz\|_F^2. \quad (46)$$

The EnKF approach then assumes that $\mu_{n|n-1}(z)$ is the Gaussian measure $N(\bar{z}_n^f, C_n^f)$ where \bar{z}_n^f and C_n^f are the ensemble mean and covariance defined in (35) and (36) respectively. Given this Gaussian assumption, it is not difficult to see that (45) implies that $\mu_{n|n}(z) = N(\bar{z}_n^a, C_n^a)$ with

$$\bar{z}_n^{(a)} = \bar{z}_n^{(f)} + K_n(y_n - H\bar{z}_n^{(f)}) \quad (47)$$

$$C_n^a = (I - K_n)C_n^f \quad (48)$$

and K_n defined in (38). In [16, Appendix A] it has been shown that the ensemble updates defined in (37) are samples from $\mu_{n|n}(z) = N(\bar{z}_n^a, C_n^a)$. In fact, [16, Appendix B] shows that the analysis step (37) can be derived from an application of RML under the Gaussian approximation $\mu_{n|n-1}(z) \approx N(\bar{z}_n^f, C_n^f)$. Indeed, it is straight forward to show that (37) can be obtained from

$$z_n^{(j,a)} = \operatorname{argmin}_z \left(\|\Gamma_n^{-\frac{1}{2}}(y_n^{(j)} - Hz)\|^2 + \|(C_n^f)^{-\frac{1}{2}}(z - z_n^{(j,f)})\|^2 \right) \quad (49)$$

which is a sequential version of (27), for the augmented state z with a prior $N(\bar{z}_n^f, C_n^f)$ and the linear measurement operator H .

For our evaluation and comparison of techniques, we consider the outcome of the EnKF algorithm after all data has been assimilated in the time interval $[0, T]$. In other words, we are interested in $\mu_{n|N_m}(z_n) = \mathbb{P}(z_n | \{y_k\}_{k=1}^{N_m})$ which corresponds to the probability of z_n after all data has been assimilated (recall N_m is the total number of assimilation times). Then, the posterior $\mu_{n|N_m}(z_n)$ computed via the EnKF algorithm provides an approximation to μ defined in (20).

4.3.3 Further Modifications

While the standard version of EnKF has been successfully applied for some history matching problems, several shortcomings due to sampling error have been identified. In particular, when a small ensemble is used, spurious correlations often cause gross over-estimation of the physical variables that EnKF aims at recovering (e.g. permeability). In addition to the issues caused by small sample size, standard EnKF with a small ensemble is suboptimal when a large amount of data are assimilated. This can be easily observed from the two following properties of EnKF. First, the ensemble updates (37), when projected into the parameter space are a linear combination of the initial ensemble members [16, 18]. Second, the ensemble updates minimize (49) which involves fitting data at each assimilation time. Therefore, when the prior ensemble is small, the EnKF updates cannot fit large amount of data within the subspace generated by the prior ensemble. These shortcomings of using standard EnKF have given rise to several EnKF variants designed to reduce the spurious correlations described above as well as increasing the number of degrees of freedom. In this work we focus on the application of distance-based covariance localization which has recently been investigated in [5, 11]. In particular, the EnKF with localization that we implement for the forward models of Section 2 is given by the same EnKF algorithm described before, except that (38) is replaced by

$$K_n = \rho \circ C_n^f H^T \left(H(\rho \circ C_n^f) H^T + \Gamma_n \right)^{-1}. \quad (50)$$

Here ρ , to be defined below, is a positive-definite matrix which induces localization and the matrix $\rho \circ C_n$ is the Schur product between ρ and C_n with entries defined by $[\rho \circ C_n]_{ij} = [\rho]_{ij} [C_n]_{ij}$. Due to the spurious correlations described above, matrix C_n^f may become positive semi-definite and the parameter update then lies in smaller subspace than the one generated by the prior ensemble. With properly chosen ρ the matrix $\rho \circ C_n^f$ has full rank, and replacing C_n^f with $\rho \circ C_n^f$ increases the dimension of the linear subspace where the parameter update is sought. This, in turn, results in a better estimation. In terms of the block structure previously described, covariance localization becomes

$$z_n^{(j,a)} = z_n^{(j,f)} + \rho_{zw} \circ C_n^{zw} (\rho_{ww} \circ C_n^{ww} + \Gamma_n)^{-1} (y_n^{(j)} - M_n(\Psi_n(v_{n-1}^{(j,a)}, u_{n-1}^{(j,a)}))) \quad (51)$$

As in the covariance localization approach of [5], we consider only localization in the u -component (e.g. for the log-permeability updates). In other words,

$$u_n^{(j,a)} = u_n^{(j,f)} + \rho_{uw} \circ C_n^{uw} (\rho_{ww} \circ C_n^{ww} + \Gamma_n)^{-1} (y_n^{(j)} - M_n(\Psi_n(v_{n-1}^{(j,a)}, u_{n-1}^{(j,a)}))) \quad (52)$$

$$v_n^{(j,a)} = v_n^{(j,f)} + C_n^{vw} (C_n^{ww} + \Gamma_n)^{-1} (y_n^{(j)} - M_n(\Psi_n(v_{n-1}^{(j,a)}, u_{n-1}^{(j,a)}))) \quad (53)$$

$$w_n^{(j,a)} = w_n^{(j,f)} + C_n^{ww} (C_n^{ww} + \Gamma_n)^{-1} (y_n^{(j)} - M_n(\Psi_n(v_{n-1}^{(j,a)}, u_{n-1}^{(j,a)}))) \quad (54)$$

Following the implementation of [11], each column of the localization matrix ρ_{uw} is defined as the fifth order compact function of Gaspari-Cohn [14] localized at the corresponding measurement location. Each row of the matrix ρ_{ww} in (52) is obtained

from ρ_{uw} by projecting it on the corresponding measurement location. By construction, ρ_{uw} and ρ_{ww} are positive definite.

Recent publications [11, 5] have investigated optimal choices for the critical length of the correlation function used for distanced-based localization. The focus of those investigations is to improve the ability of the EnKF with localization to recover the truth within the confidence interval provided by the ensemble. In contrast to [11, 5], our goal is to assess the performance of EnKF with localization for reproducing the uncertainty quantified by the posterior. However, for the present work we consider a fixed critical length obtained from a simple trial-error procedure, that enables us to observe significant effect of covariance localization in characterizing the posterior distribution. While the optimal choice of covariance localization is beyond the scope of the present work, we recognize the importance for assessing optimal choices of covariance localization for providing better Gaussian approximation of the posterior distribution at a reasonable computational cost. Moreover, additional forms of covariance regularization (e.g. covariance inflation) should also be assessed.

4.3.4 Ensemble square root filter (EnSRF)

Sampling error that arises from perturbing the observations in standard EnKF has been often associated with a poor performance of history matching data. In order to avoid the aforementioned sampling error, an ensemble square root filter (EnSRF) is often used. Here we consider the following EnSRF [10]:

Algorithm 5 (EnSRF) Construct an initial ensemble as in (33). For $j = 1, \dots, N_m$

- (1) *Prediction Step:* Propagate the ensemble of particles forward under (31) yielding (34). Construct the sample mean and covariance from (35)-(36). Additionally define the deviations from the mean

$$\Delta z_n^{(j,f)} = z_n^{(j,f)} - \bar{z}_n^f. \quad (55)$$

- (2) *Analysis step:* Compute the updated mean $\bar{z}_n^{(a)}$ via formula (47) with K_n given by (38). Consider the matrices $\Delta Z_n^f := [\Delta z_n^{(1,f)} \ \Delta z_n^{(2,f)} \ \dots \ \Delta z_n^{(N_e,f)}]$, with j th column $\Delta z_n^{(j,f)}$, and ΔZ_n^a defined analogously. Compute the matrix with updated deviations,

$$\Delta Z_n^a = (I - \tilde{K}_n H) \Delta Z_n^{(f)} \Theta \quad (56)$$

with

$$\tilde{K}_n = C_n^f H^T \left[H_n C_n^f H_n^T + \Gamma_n \right]^{-T/2} \left[(H_n C_n^f H_n^T + \Gamma_n)^{1/2} + \Gamma_n^{1/2} \right]^{-1} \quad (57)$$

The updated ensemble is then obtained from the expression

$$z_n^{(j,a)} = \bar{z}_n^{(a)} + \Delta z_n^{(j,a)}. \quad (58)$$

In expression (56), Θ is a $N_e \times N_e$ mean-preserving orthogonal random matrix constructed as suggested in [23]. The mean-preserving property of Θ ensures that $\sum_{j=1}^{N_e} \Delta z_n^{(j,a)} = 0$ and so the analyzed ensemble (58) has mean $\bar{z}_n^{(a)}$ as required. In contrast to the EnKF, where (48) is only exactly satisfied in the limit of arbitrarily large ensemble size, the sample covariance computed from the EnSRF (finite) ensemble updates exactly satisfy (48), therefore providing a better approximation of $\mu_{n|n}(z)$; it is then hoped that this will lead to a better approximation to the posterior distribution (20) itself. A block structure similar to the one introduced before applies to the EnSRF. From this structure it is easy to appreciate that only small matrices are involved in the square root computations. Therefore, the computational cost of EnSRF is essentially the same as EnKF, i.e., N_e times the number of forward model evaluations.

Even though the implementation of the EnSRF avoids the sampling error due to perturbing the observations, limitations related to the small ensemble size still apply. Distance-based localization can then be applied to the EnSRF as suggested in [10]. Concretely, we replace K_n in (38) with (50) and \tilde{K}_n in (55) with

$$\tilde{K}_n = (\rho \circ C_n^f) H^T \left[H_n (\rho \circ C_n^f) H_n^T + \Gamma_n \right]^{-T/2} \left[(H_n (\rho \circ C_n^f) H_n^T + \Gamma_n)^{1/2} + \Gamma_n^{1/2} \right]^{-1} \quad (59)$$

Similar to the localization procedure for the EnKF, EnSRF is localized only in the u -component (log-permeability) of (47) and (54) with the localization matrix ρ described above.

5 Numerical Results

In this section we present the results of three numerical experiments for assessing the Gaussian approximations defined in Section 4. These experiments are described on the three subsections which follow, each of which is organized as follows: (i) details of the forward model and the generation of synthetic data are provided; (ii) numerical results from the gold-standard MCMC implementation described in Section 3 are discussed; (iii) the numerical results of Gaussian approximations are presented and the results in (iii) are compared against the results from (ii) in terms of their ability to reproduce mean and variance; (iv) we assess the performance of the Gaussian approximation at quantifying the uncertainty in reservoir model forecast.

5.1 Single-Phase flow

For the first experiment we consider the single-phase reservoir model of Section 2 on a square domain $D = [0, L] \times [0, L]$ with the production wells located at the points labeled by P_1, \dots, P_9 in Figure 1 (middle). Relevant information of this model is displayed in the first column of Table 1. For each well term in the right hand side of (1) we prescribe a production rate of $85\text{m}^3/\text{day}$ constant during the total simulation time of 50 days.

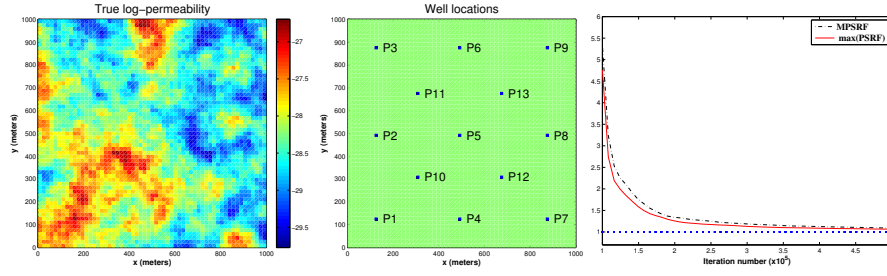


Fig. 1 Single-phase model. Left: True log-permeability [$\log \text{m}^2$]. Middle: Well configuration. Right: Gelman-Rubin diagnostic

We consider a Gaussian prior distribution of log-permeability

$$\mu_0(u) = N(\bar{u}, C) \quad (60)$$

where the covariance is defined by $C = \kappa A^{-\alpha}$, with the operator $A = -\Delta$ defined on

$$D(A) = \{v \in H^2(D) | \nabla v \cdot \mathbf{n} = 0 \text{ on } \partial D, \int_D v = 0\} \quad (61)$$

i.e. A is the negative Laplacian with no-flow boundary conditions and restricted to spatial average zero functions. The tunable parameters in (60) are defined as follows: $\bar{u}(x, y) = \log(5 \times 10^{13} \text{m}^2)$ for all $(x, y) \in D$, $\kappa = 2.0$ and $\alpha = 1.3$. In Figure 2 we show some realizations of the prior distribution (60). It is important to mention that other choices of C can also be used. In particular, C can be defined in terms of a standard correlation function (e.g. spherical, exponential, etc). Our choice, however, has the advantage that C becomes a diagonal operator in the spectral domain. Sampling from the prior on the spectral domain is straightforward and computationally inexpensive. This is a desirable property since at each iteration of MCMC (see equation (22)), a draw from the prior is generated for computing the proposal. The correlation function of draws from the prior is, of course, simply the Green's function of C .

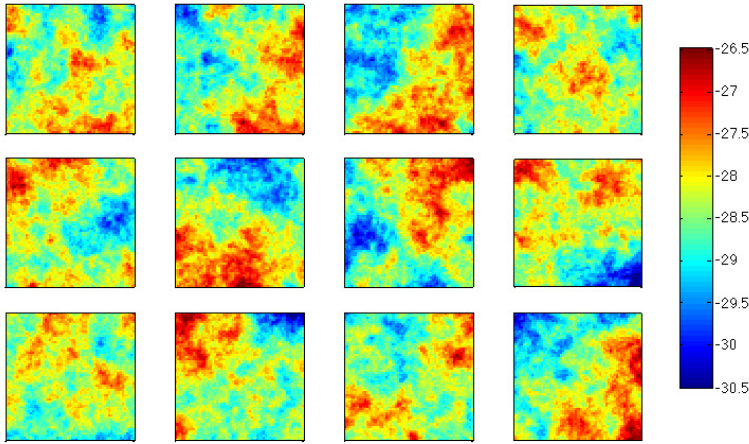
For the generation of synthetic pressure data, we first define the “true log-permeability” denoted by u^\dagger and displayed in Figure 1 (left). This “true log-permeability” is generated from the prior distribution defined above. Synthetic data is generated by first solving (1)-(3) for p with $u = u^\dagger$. Then, $G(u^\dagger)$ is calculated from (6). Finally, we add random error, i.e. $y = G(u^\dagger) + \eta$ with $\eta \sim N(0, \sigma^2 I)$ with $\sigma = 4 \times 10^5 \text{Pa}$. The measurement times used in (4) are $t_1 = 5$, $t_n = 10n$ days, $n = \{2, \dots, 5\}$.

Synthetic data are used in the pCN-MCMC Algorithm 1 with $\beta = 0.015$. Our MCMC results consist of 110 chains starting from independent draws from the prior distribution. After a burn-in period of 1×10^4 , each chain generates 5×10^5 samples. For assessing the convergence of our chains, we consider the diagnostics suggested by Gelman and Rubin in [4]. In Figure 1 (right) we display the maximum of the potential scale reduction factor (PSRF) and the multivariate potential scale reduction factor (MPSRF) for the smallest $J = 16$ frequencies that account for 76% of the total prior energy defined by $e(J) = \sum_{j=1}^J \lambda_k / \sum_{j=1}^\infty \lambda_k$ where λ_k are the eigenvalues of the prior covariance C (ordered as $\lambda_1 \geq \lambda_2 \geq \dots$). Convergence of the chains is achieved when

Table 1 Reservoir model description

Variable	single-phase reservoir	water-oil reservoir (small number of wells)	water-oil reservoir (large number of wells)
L [m^3]	10^3	2×10^3	5×10^3
c [Pa^{-1}]	10^{-8}	0.0	0.0
v_o [Pa s]	10^{-2}	10^{-2}	10^{-2}
T [years]	0.13	5	3.5
^a p_0 [Pa]	3.5×10^7	2.5×10^7	2.5×10^7
^a s_0	not applicable	0.2	0.2
v_w [Pa s]	not applicable	5×10^{-4}	5×10^{-4}
s_{iw}	not applicable	0.2	0.2
s_{ro}	not applicable	0.2	0.2
^b p_{ph}^i [Pa]	not applicable	2.7×10^7	2.0×10^7
^b q_w^i [m^3/day]	not applicable	2.6×10^3	1.8×10^2
a_w	not applicable	0.3	0.3
a_o	not applicable	0.9	0.9

^a Constant in Ω . ^b Constant in $[0, T]$.

**Fig. 2** Single-phase model. Samples from the prior distribution [$\log \text{m}^2$].

the maximum of the PSRF and the MPSRF are close to one. From Figure 1 (right), the $\max(\text{PSRF})$ and the MPSRF have dropped below 1.1 after 5×10^5 iterations where we establish the convergence of our MCMC chains. From the numerical evidence of convergence of our chains, we conclude that the MCMC provides samples from the posterior. The associated mean and variance fields, denoted by $u_{pos}(x)$ and $\sigma_{pos}(x)$, are used as gold standard for the assessment of the Gaussian approximations that we discuss below. In Figure 3 we display some samples from the independent chains (i.e. uncorrelated) obtained after convergence was achieved. Although there are substantial differences among those realizations, some common spatial features can be observed. For example, note the high permeability region around wells P_1 and P_6 .

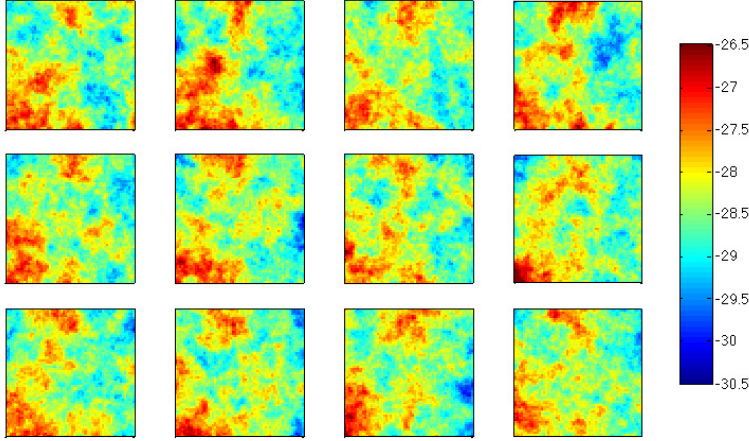


Fig. 3 Single-phase model. Samples from the posterior distribution (characterized with MCMC) [$\log m^2$]

Furthermore, the variability is considerably lower than under the prior, as exhibited in Figure 2; this indicates that the data used is quite informative.

The numerical implementation of the following Gaussian approximations are conducted for an ensemble of size $N_e = 50$: LMAP, RML, EnKF, EnKF with localization, EnSRF and EnSRF with localization. It is important to emphasize that the mean and variance for (the sequential methods) EnKF, EnKF with localization, EnSRF and EnSRF with localization, are computed after all the measurements have been assimilated. In other words, exactly the same data used to sample the posterior via MCMC are also used for all the Gaussian approximations under consideration. From each of these approximations, we compute the mean and variance which we compare against the mean u_{pos} and variance σ_{pos} of the posterior distribution generated with the MCMC method. The mean and variance are shown in Figure 4 and Figure 5, respectively. In Table 2 we display the relative errors of the deviation of the mean with respect to \bar{u} (the prior mean) and the relative error of the variance defined by

$$\epsilon_u = \frac{\|(\hat{u} - \bar{u}) - (u_{pos} - \bar{u})\|_{L^2(D)}}{\|(u_{pos} - \bar{u})\|_{L^2(D)}}, \quad \text{and} \quad \epsilon_\sigma = \frac{\|\hat{\sigma} - \sigma_{pos}\|_{L^2(D)}}{\|\sigma_{pos}\|_{L^2(D)}}, \quad (62)$$

respectively. In the previous expression, \hat{u} and $\hat{\sigma}$ are the mean and variance of the Gaussian approximation under consideration. The right column of Table 2 indicates the computational cost for computing each of the techniques in terms of forward model runs. As stated in Section 4, while the computational cost of the Kalman filter-type of methods is stable with respect to the details of the implementation, the cost of RML and LMAP depends crucially on the optimization technique used for solving (24) and (27). Our implementation of the Levenberg-Marquardt technique cost around 5 forward model runs per iteration. Furthermore, in average, each of optimization problems (27) converged in 5 iterations and so the average computational

cost of (27) is 25 forward model runs. This computational cost can be potentially reduced by applying a more efficient optimization technique.

Since we are assessing the Gaussian approximations only in terms of mean and variance, we can additionally measure the error (with respect to the posterior) of the exact mean and variance of $N(u_{MAP}, C_{MAP})$ given directly (24) and (25), respectively. The corresponding relative errors are provided in the “MAP” row in Table 2. From construction it is clear that the mean and covariance of LMAP are u_{MAP} and C_{MAP} for sufficiently large N_e . The results for this experiment indicate that $N(u_{MAP}, C_{MAP})$ provides a good approximation to the posterior in terms of mean and variance. Therefore, more samples from LMAP can be generated at a negligible cost so that its mean and variance approaches u_{MAP} and C_{MAP} , respectively.

Table 2 also indicates that, among the all the Gaussian approximation with $N_e = 50$, RML provides the best approximation in terms of the mean. The worst performance in terms of mean and variance was obtained with EnKF. However, considerable improvement was obtained by applying the localization approach described in the preceding section (see equation (52)). We recall from exposition of Section 4 that EnSRF reduces the sampling error that arises from standard implementations EnKF where data are perturbed with noise. From Table 2 we observe that the effect of sampling error has a detrimental effect in the performance of EnKF for reproducing the posterior distribution. More precisely, EnSRF outperformed the EnKF both in terms of mean and variance with respect to the posterior distribution. Note that the application of localization in the EnSRF (expression (55)) further reduces the relative errors in the mean and variance. In fact, among all techniques with $N_e = 50$, the best approximation in terms of variance is given by the EnSRF with localization.

From the preceding comments it clear that sampling error causes severe limitations in the performance of EnKF and EnSRF. It is worth mentioning that issues of the EnKF and EnSRF due to sampling error have been often reported [5, 21] and used as motivation for covariance regularization. However, this existing work is focused on (i) history matching production data (ii) recovering the true permeability and (iii) recovering data generated with the true permeability within the estimated confidence interval. In contrast, here we assess the performance in terms of the posterior distribution of the Bayesian framework.

Since sampling error due to the small ensemble size severely limits the ability of EnKF to produce reasonable approximations of the posterior, we consider three more additional implementations of EnKF for larger ensembles: $N_e = 1250$, $N_e = 2500$ and $N_e = 5000$. These results appear at the end of Table 2. Note that $N_e = 1250$ corresponds to the case where the computational cost of EnKF coincides with the cost of our implementation of RML. While the performance of EnKF improved significantly for $N_e = 1250$, RML still provides a better approximation in terms of the mean. In addition, we observe that although increasing the size of the ensemble may reduce the sampling error, this is not associated to the convergence to the posterior. Actually, it is clear from this experiment that the variance of EnKF for large N_e seems to diverge from the variance of the posterior.

With the previous results we are able to appreciate and evaluate the differences in the approximations provided by each of the Gaussian approximations of the posterior. This posterior is the conditional probability of the unknown (permeability) given

production data collected during the 50 days of simulation. For practical applications it is of particular interest to assess how different Gaussian approximations fare at reproducing the probability distribution of various predicted quantities, with respect to the posterior; in other words to assess how the approximate algorithms fare in the quantification of uncertainty in these predictions. The assessment of performance in terms of the distribution of predictions under the posterior is conducted by creating a new flow scenario as we now describe. Assume that after the initial 50 days of simulation (that we used to generate synthetic data), we now drill new wells labeled by P_{10} , P_{11} , P_{12} and P_{13} in Figure 1. These new wells are operated at constant production rate of $60\text{m}^3/\text{day}$ during 100 days. During this 100 days of forecast, the old wells P_1, \dots, P_9 are first shut-down for a pressure build-up time window of 50 days, followed by a constant production of $60\text{m}^3/\text{day}$ during the rest 50 days. In Figure 6 we show, as a function of time, the pressure at the well locations P_9 , P_{10} and P_{13} . The first 50 days corresponds to the data assimilation phase and the subsequent 100 days are the prediction. In the first row of Figure 6 we display the pressure obtained from the reservoir simulation with 100 permeabilities obtained from the prior distribution (60). The second row corresponds to the pressure obtained from the simulation with the samples from the posterior obtained from independent MCMC chains. Subsequent rows of Figure 6 corresponds to pressure obtained from simulating the permeabilities obtained from some of the Gaussian approximations under consideration. The vertical line divides data assimilation phase from the forecast. Additionally, since we are interested in the performance with respect to the posterior, for each curve presented in Figure 6 we include a red curve of the pressure at the corresponding well location obtained by simulating the mean of the posterior distribution (i.e. top-left field of Figure 4).

Note that the uncertainty quantified by the posterior and the Gaussian approximations is considerably small at the wells P_1, \dots, P_9 where measurements were collected. In contrast, large uncertainty in the forecast is observed for the new wells P_{10}, \dots, P_{13} for which data was not available during the data assimilation phase. Note for example that in P_{11} , the posterior is visually close to the prior, indicating the uninformative effect of the data at the location of P_{11} . For the new wells where uncertainty is larger, we can appreciate the performance on the Gaussian approximations. Note from Figure 6 that the EnKF (without localization) at well P_{11} underestimates the uncertainty in the model predictions.

In Figure 7 we display the distribution of the pressure at the new wells P_{10}, \dots, P_{13} at the final time $t = 150$ days. The horizontal line correspond to the value of the pressure at the corresponding location obtained from simulating the model with the posterior mean. From 6 and 7 we observe that LMAP and RML provide a better approximation to the predicting distribution than the EnKF-based methods. Since RML produces the best approximation of the posterior in terms of mean with a reasonable approximation of the variance, the associated approximation of the predicting distribution is the most accurate among the techniques considered here. Our results with respect to the optimality of RML for this experiment are similar to those reported in [17] where a single-phase reservoir model was also utilized. As we will see in the next experiments, a less favorable performance of RML is observed for a two-phase reservoir model.

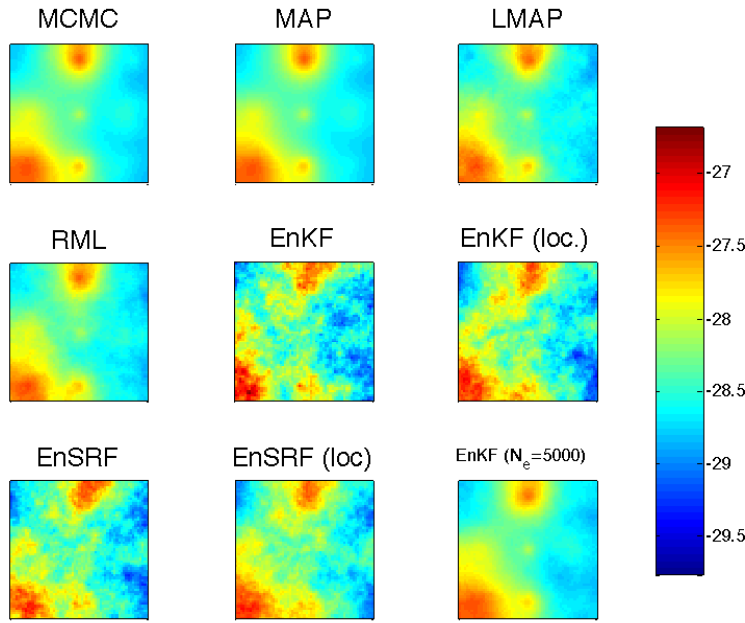


Fig. 4 Single-phase model. Mean of the posterior distribution (characterized with MCMC) and Gaussian approximations [$\log \text{m}^2$]

Table 2 Evaluation of Gaussian approximations on the single-phase model

Method	Relative error in the mean ε_μ	Relative error in the variance ε_σ	Computational cost [Forward model runs]
MCMC	0.000	0.000	5.5×10^7
MAP	0.030	0.094	2.5×10^1
LMAP ($N_e = 50$)	0.179	0.259	2.5×10^1
RML ($N_e = 50$)	0.154	0.258	1.25×10^3
EnKF ($N_e = 50$)	0.643	0.417	5.0×10^1
EnKF (localization, $N_e = 50$)	0.546	0.288	5.0×10^1
EnSRF ($N_e = 50$)	0.519	0.267	5.0×10^1
EnSRF (localization, $N_e = 50$)	0.445	0.208	5.0×10^1
EnKF ($N_e = 1250$)	0.192	0.075	1.25×10^3
EnKF ($N_e = 2500$)	0.120	0.089	2.5×10^3
EnKF ($N_e = 5000$)	0.102	0.094	5.0×10^3

5.2 Oil-water reservoir: Small number of wells

In this subsection we consider the oil-water reservoir model described in Section 2. The reservoir is a square $D = [0, L] \times [0, L]$ with a five-spot well configuration consisting on production wells P_1, \dots, P_4 and one injection well I_1 as displayed in Figure 8, middle (well P_5 will play a role in later discussions). Relevant information concerning this model is displayed in Table 1. The prior distribution of log-permeability

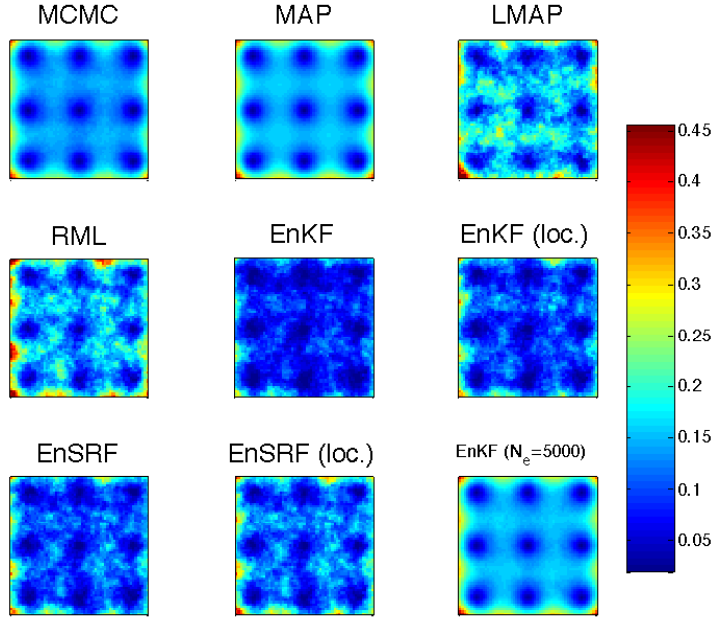


Fig. 5 Single-phase model. Variance of the posterior distribution (characterized with MCMC) and Gaussian approximations $[(\log m^2)^2]$

is defined by the Gaussian measure defined in (60) with same parameters except κ which for this experiment is $\kappa = 4.0$. With these parameters, similar realizations to those of Figure 2 are obtained but with a range of variability with respect to the prior mean \bar{u} increased by a factor of $\sqrt{2}$.

Similarly to the previous example, for the generation of synthetic data we define the “true log-permeability” u^\dagger displayed in Figure 8 (left). This “true log-permeability” is generated from the prior distribution described in the preceding paragraph. Note that u^\dagger is the same as the one used for the previous experiment (see Figure 1 (left)) but with the magnitude of $u^\dagger - \bar{u}$ multiplied by $\sqrt{2}$. The generation of synthetic data is now conducted by computing (p, s) from (7)-(8) with $u = u^\dagger$. Equation (17) is then used to compute $G(u^\dagger)$ and zero mean Gaussian random error η is added to generate data $y = G(u^\dagger) + \eta$. The measurement times used in (14)-(15) are $t_n = 0.67n$ years, $n = \{1, \dots, 7\}$. According to the structure of (16)-(17), η has the following form

$$\eta = (\eta_1, \dots, \eta_{N_M}), \quad \eta_n = (\eta_n^{1,I}, \dots, \eta_n^{N_I,I}, \eta_n^{1,P}, \dots, \eta_n^{N_P,P}) \quad (63)$$

We generate the components of η as follows

$$\begin{aligned} \eta_n^{1,I} &\sim N(0, (3.2 \times 10^4 \text{Pa})^2) & \eta_n^{1,P} &\sim N(0, (0.25 \text{m}^3/\text{day})^2) \\ \eta_n^{2,P}, \eta_n^{3,P}, \eta_n^{4,P} &\sim N(0, (0.02 \text{m}^3/\text{day})^2). \end{aligned} \quad (64)$$

for all $n \in \{1, \dots, 7\}$. In the previous definitions we consider larger measurement error at the production well P_1 since larger total flow rates are obtained. This is caused by

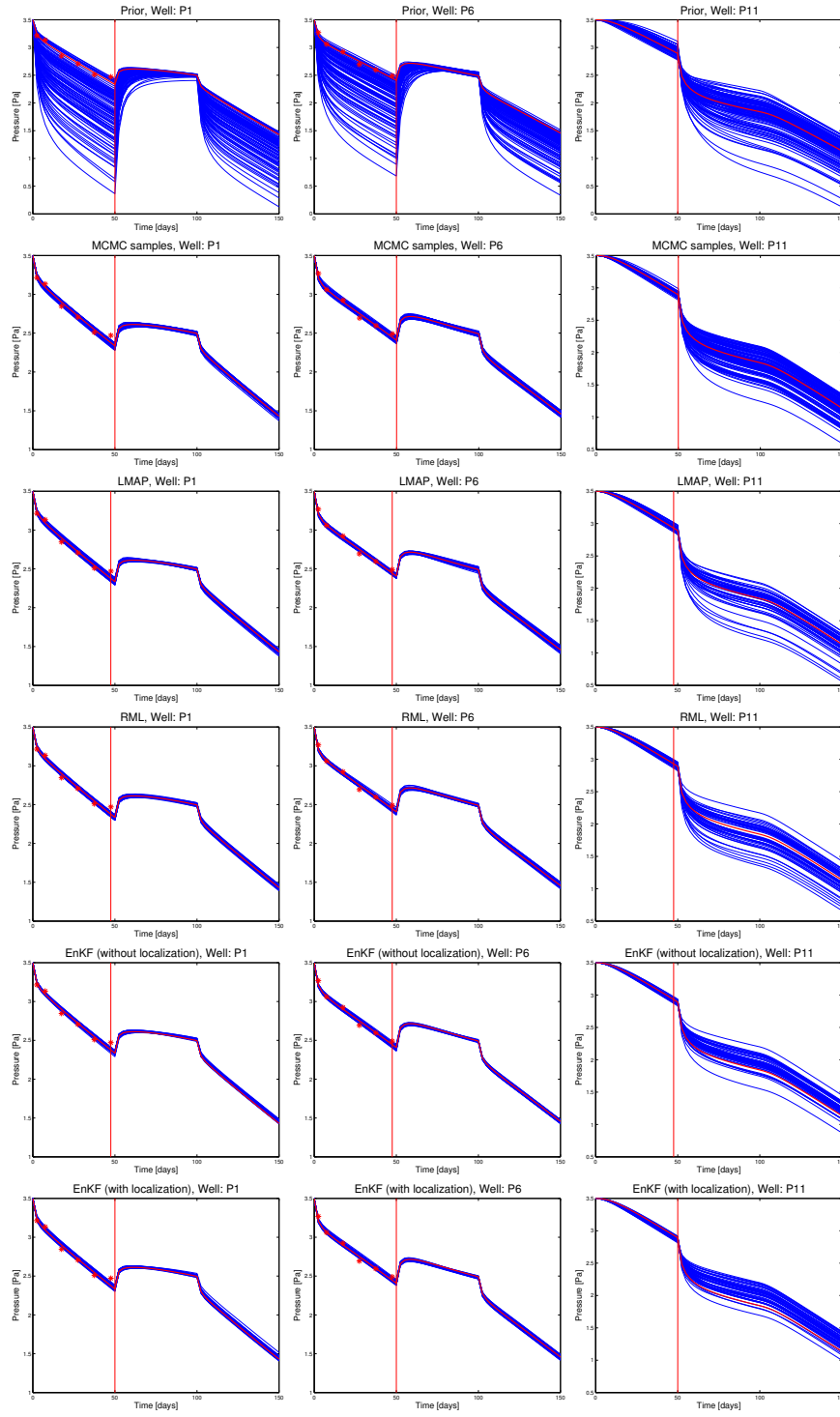


Fig. 6 Single-phase model. Pressure from wells P_{11} (right column), P_6 (middle column) and P_1 (left column) simulated with permeabilities sampled from (top to bottom rows) the prior, the posterior, LMAP, RML, EnKF and EnKF with localization.

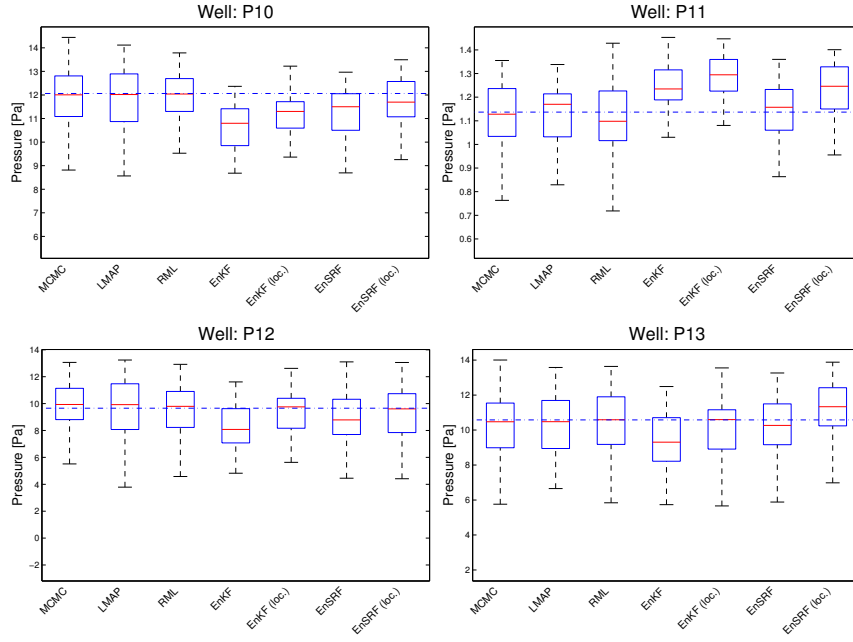


Fig. 7 Single-phase model. Distribution of pressure at wells P_{10}, \dots, P_{13} at the final time of simulation $t = 150$ days

the larger permeability region around P_1 and so the early water breakthrough at this well.

The pCN-MCMC Algorithm 1 with $\beta = 0.015$ is applied to this problem with the synthetic data previously described. 110 chains are generated starting from independent draws from the prior distribution. A burn-in period of 1.5×10^4 was observed after which the chains were run 5×10^5 iterations. By using the same Gelman-Rubin indicated in the previous example, we determine convergence of our chains. This numerical evidence is presented in Figure 8 where, after 5×10^5 iteration, both the max PRSF and MPRSF for the highest energy models converges to one. Uncorrelated samples (from independent chains) are shown in Figure 9.

Samples of the posterior from our converged chains define our gold standard. These are then used to compare the performance of Gaussian approximations in terms of mean and variance. Analogous to the previous example, we use an ensemble with of size $N_e = 50$ and compute the mean and variance with LMAP, RML, EnKF, EnKF with localization, EnSRF and EnSRF with localization. Relative errors of the mean and variance (see expression (62)) with respect to the posterior distribution are provided in Table 3. In Figure 10 and Figure 11 we present the mean and variance, respectively.

In contrast to the previous experiment, the error in the approximation given by $N(u_{MAP}, C_{MAP})$ is relatively large. The relative errors of mean and variance are 27% and 14% respectively. Similar to the previous experiment, RML provided the best approximation of the posterior in terms of the mean. However, the variance of the

posterior was significantly overestimated by RML. The performance of the standard EnKF for $N_{en} = 50$ was very poor. From Figure 10 we observe large values of the estimated field, which are typically found in standard EnKF applications with small ensemble sizes. Nevertheless, for the same size, covariance localization has a positive effect by reducing the error in the mean and variance with respect to the posterior. Similar to the previous experiment, EnKF and EnKF with localization were outperformed by the corresponding EnSRF implementations. As in the previous experiment, EnSRF with localization provided, among the techniques with $N_e = 50$, the best approximation in terms of variance.

Unlike the preceding experiment, here we observe that increasing the size of the ensemble does not result in a decrease of the error with respect to the mean. This can be observed from at the end of Table 3 where we report the results of EnKF implementations for $N_{en} = 1000$, $N_{en} = 3000$, $N_{en} = 8000$. Both in terms of mean and variance, EnKF for large ensembles does not exhibit convergence to the posterior. Note that EnKF with $N_{en} = 3000$ corresponds to the same computational cost of RML. Yet, the latter provides a better approximation in terms of the mean. Among all the techniques, LMAP provides reasonable approximations of both the mean and variance of the posterior.

In order to assess the performance of the approximate posterior samples at reproducing the predicting distribution, we consider an additional simulation period of 5 years of forecast. For this additional 5 years, a new well labeled as P_5 in Figure 8 (middle) is drilled and operated under constant fixed bottom-hole pressure $P_{bh}^5 = 2.7 \times 10^7$ Pa. In Figure 12 we present the total flow rates (from P_1 and P_5) and bottom-hole pressure (from I_1) simulated with permeabilities from the prior (first row), the posterior (second row), and some of the Gaussian approximations under analysis (third-sixth row). The vertical line divides the assimilation from the prediction. The red curve is computed from the posterior mean at the corresponding location. We note that the poor performance of EnKF is reflected in the model predictions whose ensemble does not capture the prediction based on the mean of the posterior. These issues are alleviated by localization as we observe from Figure 12 (last row).

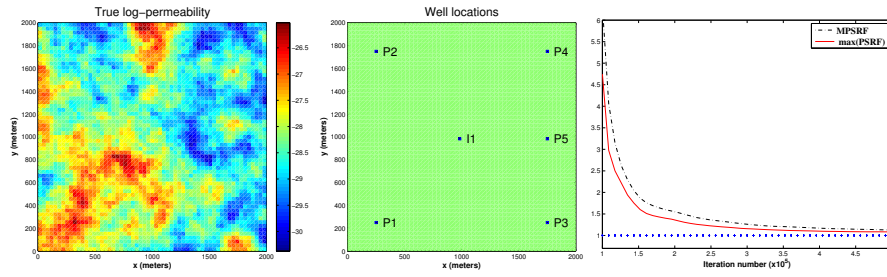
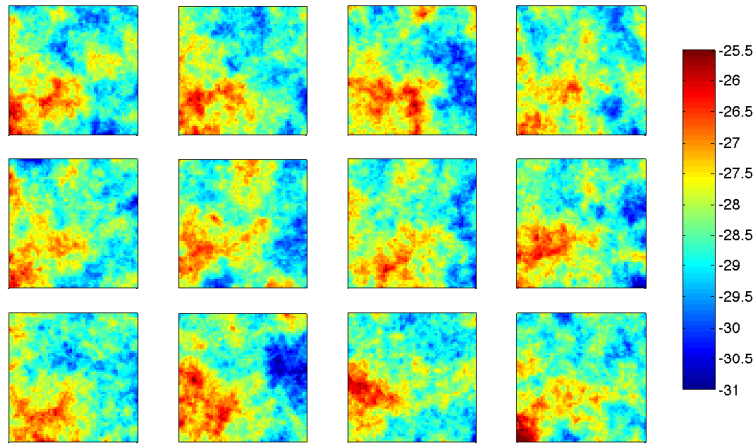
In Figure 13 we display the distribution of the final time cumulative oil production simulated from the posterior and the Gaussian approximation. Note that even though LMAP provides a reasonable approximation in terms of both mean and variance, the approximation provides a deficient characterization of the predicting distribution. In general all the Gaussian approximations exhibit poor performance at reproducing the predicting distribution.

5.3 Oil-water reservoir: Large number of wells

In this subsection we consider again the oil-water reservoir model described in Section 2. The well configuration for this case is displayed in 14 (middle) and relevant information can be found in Table 1. The aim of this experiment is to evaluate the performance of Gaussian approximations when measurements from many wells are available. The prior distribution of log-permeability is defined by the Gaussian mea-

Table 3 Evaluation of Gaussian approximations for the two-phase model. Case with small number of wells.

Method	Relative error in the mean ε_μ	Relative error in the variance ε_σ	Computational cost [Forward model runs]
MCMC	0.000	0.000	5.5×10^7
MAP	0.277	0.143	6.0×10^1
LMAP ($N_e = 50$)	0.284	0.286	6.0×10^1
RML ($N_e = 50$)	0.253	0.475	3.0×10^3
EnKF ($N_e = 50$)	1.159	0.424	5.0×10^1
EnKF (localization, $N_e = 50$)	0.600	0.263	5.0×10^1
EnSRF ($N_e = 50$)	0.715	0.397	5.0×10^1
EnSRF (localization, $N_e = 50$)	0.483	0.259	5.0×10^1
EnKF ($N_e = 1000$)	0.353	0.209	1.0×10^3
EnKF ($N_e = 3000$)	0.301	0.222	3.0×10^3
EnKF ($N_e = 8000$)	0.337	0.216	8.0×10^3

**Fig. 8** Two-phase model (small number of wells). Left: True log-permeability [$\log m^2$]. Middle: Well configuration. Right: Gelman-rubin diagnostic**Fig. 9** Two-phase model (small number of wells). Samples from the posterior distribution (characterized with MCMC) [$\log m^2$]

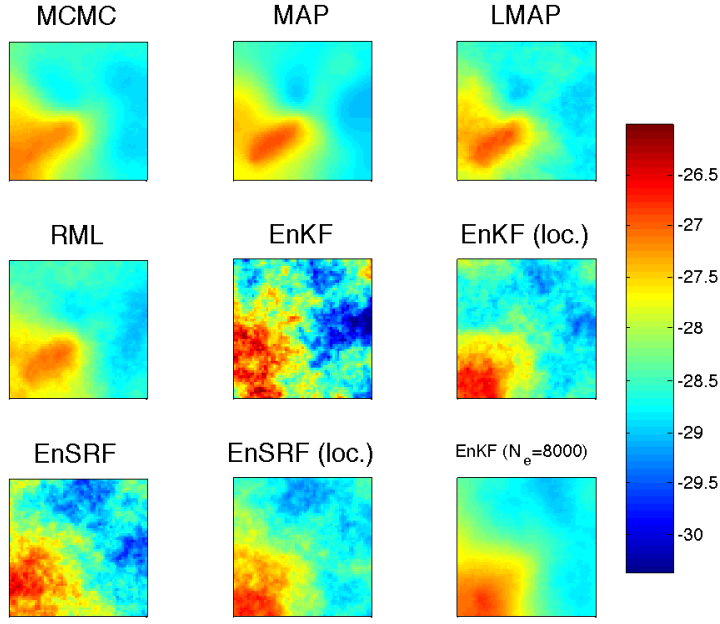


Fig. 10 Two-phase model (small number of wells). Mean of the posterior distribution (characterized with MCMC) and Gaussian approximations [$\log m^2$]

sure defined in (60) with the same tunable parameters used in the experiment of subsection 5.1.

Generation of synthetic data was conducted with the same procedure described before. The “true log-permeability” u^\dagger is displayed in Figure 14 (left). The random error η added to $G(u^\dagger)$ has the form of (63) with $N_I = 9$ and $N_P = 16$ and

$$\begin{aligned} \eta_n^{j,I} &\sim N(0, 2.7 \times 10^4 \text{Pa}), & j \in \{1, \dots, N_I\}, \\ \eta_n^{j,P} &\sim N(0, 0.06 \text{m}^3/\text{day}), & j \in \{1, \dots, N_P\}. \end{aligned} \quad (65)$$

for all $n \in \{1, \dots, 7\}$. Measurement times are $t_n = 0.467n$ years, $n = \{1, \dots, 7\}$. The pCN-MCMC Algorithm 1 is applied to generate 110 chains starting from independent draws from the prior distribution. After a burn-in period of 1.5×10^4 the chains are run for 5×10^5 iterations. The Gelman-Rubin diagnostic is conducted as describe in the preceding sections. Figure 14 (right) show the PRSF and MPRSF as defined previously. Uncorrelated samples (from independent chains) are shown in Figure 15.

Similar to the previous section, converged chains provide the posterior against which to compare the performance of Gaussian approximations in terms of mean and variance. The first part of Table 4 provides the results when an ensemble with of size $N_e = 50$ is used. In Figure 16 and Figure 17 we display mean and variance, respectively.

Among all the ensemble methods with $N_e = 50$, RML provides the best approximation in terms of mean. Note that the approximation provided by $N(u_{MAP}, C_{MAP})$

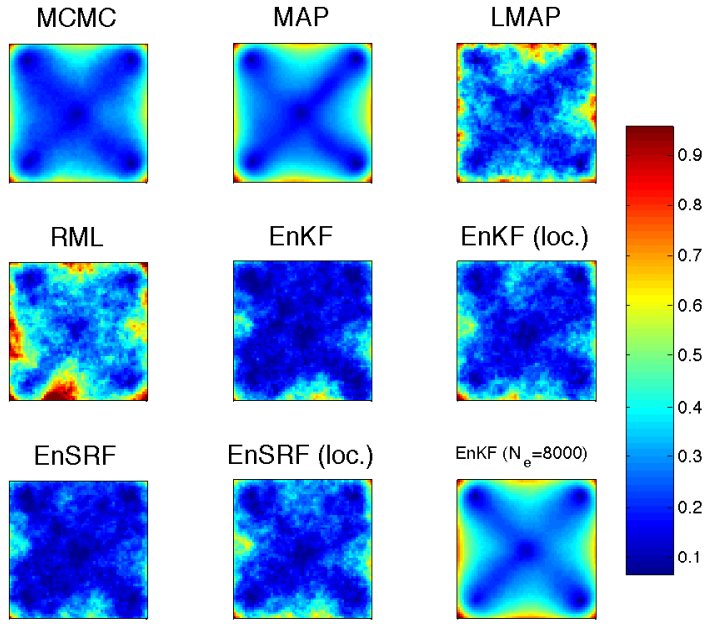


Fig. 11 Two-phase model (small number of wells). Variance of the posterior distribution (characterized with MCMC) and Gaussian approximations $[(\log m^2)^2]$

provides the best approximation in terms of combined mean and variance. Additionally, even with localization both EnKF and EnSRF provide a very poor approximation in terms of mean and variance. It is worth mentioning that RML and LMAP provided a better approximation (in terms of mean and variance) than the ensemble Kalman filter-type methods for $N_e = 50$. In Figure 18 we show the total flow rates (from P_1 and P_5) and bottom-hole pressure (from I_1) simulated with permeabilities from the prior (first row), the posterior (second row), and some of the Gaussian approximations under analysis (third-sixth row). The vertical line divides the assimilation from the prediction. In this case, prediction is performed by simulating an additional 3.5 years under the same well configuration. The red curve is computed from the posterior mean at the corresponding location. In Figure 19 we display the distribution of the final time cumulative oil production simulated from the posterior and the Gaussian approximation. The poor performance of EnKF and EnSRF is reflected in the poor performance at characterizing the predicting distribution.

As we mentioned earlier, limitations of the EnKF and EnSRF arise when large number of measurements are assimilated. In the present work we are interested in the associated detrimental effect on the approximation of the posterior distribution. In order to observe that effect, we now consider application of our Gaussian approximation on a larger ensemble $N_e = 250$. These results are presented in the second part of Table 4. Note that for $N_e = 250$ the ratio of total number of wells to ensemble size is the same as in the previous experiment ($N_e = 50$ and $N_w = 5$). For $N_{en} = 250$, Table

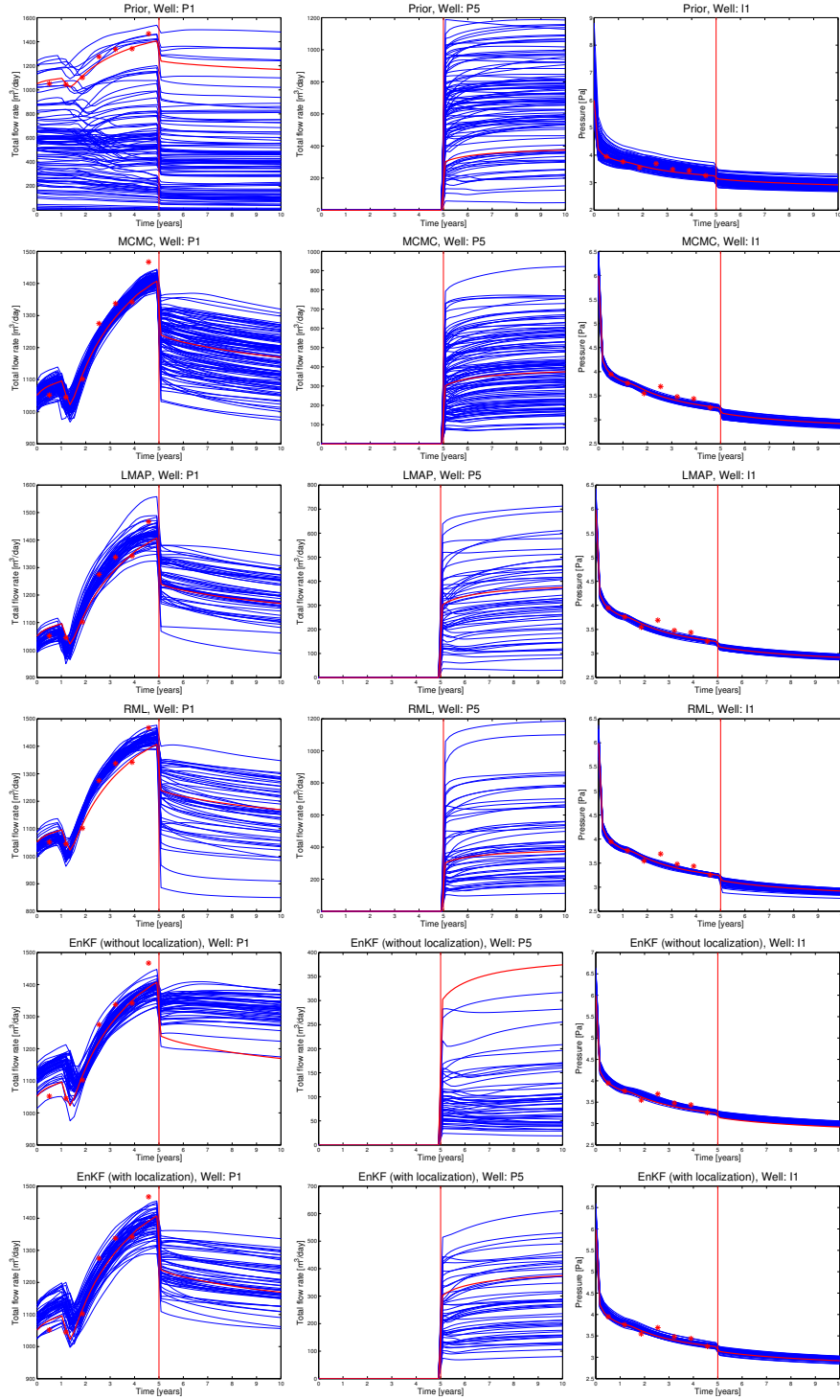


Fig. 12 Two-phase model (small number of wells). Total flow rates from P_1 (left column), P_5 (middle column) and bottom-hole pressure from I_1 (right column) simulated with permeabilities sampled from (top to bottom rows) the prior, the posterior, LMAP, RML, EnKF and EnKF with localization.

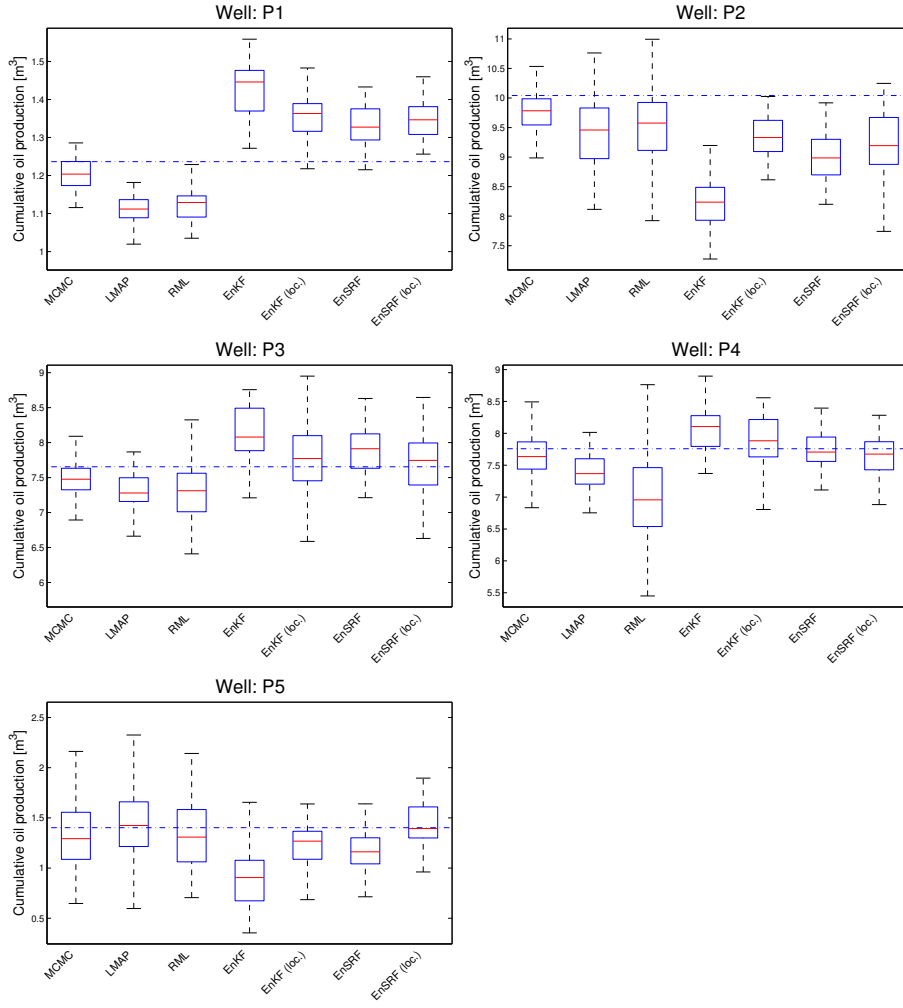
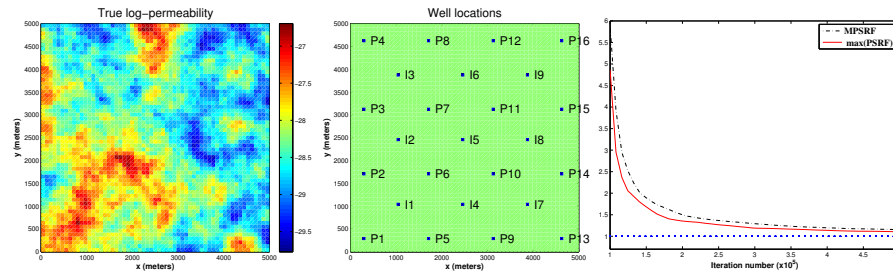


Fig. 13 Two-phase model (small number of wells). Distribution of cumulative oil production at wells P_1, \dots, P_5 at the final time of simulation $t = 10$ years

4 indicates that RML provides again the best approximation in terms of mean. The performance of EnKF, EnSRF and their localizations are considerably improved with respect to the ones for $N_e = 50$. As in previous examples, EnSRF with localization provides the best approximation in terms of variance. Also similarly to the previous experiments, increasing the size of the standard EnKF does not improve the approximation in terms of variance. On the other hand, in this case the variance increases with the size of ensemble. This can be observed from the last part of Table 4 where EnKF was implemented for $N_{en} = 1000$, $N_{en} = 3000$, $N_{en} = 6000$ and $N_{en} = 18500$. Note that the computational cost of EnKF for $N_{en} = 18500$ coincides with the cost of our implementation of RML with $N_{en} = 50$.

Table 4 Evaluation of Gaussian approximations for the two-phase model. Case with large number of wells.

Method	Relative error in the mean ε_μ	Relative error in the variance ε_σ	Computational cost [Forward model runs]
MCMC	0.000	0.000	5.5×10^7
MAP	0.131	0.165	3.5×10^2
LMAP ($N_e = 50$)	0.179	0.287	3.5×10^2
RML ($N_e = 50$)	0.169	0.307	1.85×10^4
EnKF ($N_e = 50$)	0.932	0.816	5.0×10^1
EnKF (localization, $N_e = 50$)	0.635	0.616	5.0×10^1
EnSRF ($N_e = 50$)	0.862	0.658	5.0×10^1
EnSRF (localization, $N_e = 50$)	0.539	0.471	5.0×10^1
LMAP ($N_e = 250$)	0.146	0.190	3.5×10^2
RML ($N_e = 250$)	0.121	0.231	9.25×10^4
EnKF ($N_e = 250$)	0.434	0.166	2.5×10^2
EnKF (localization, $N_e = 250$)	0.304	0.113	2.5×10^1
EnSRF ($N_e = 250$)	0.371	0.110	2.5×10^2
EnSRF (localization, $N_e = 250$)	0.285	0.101	2.5×10^2
EnKF ($N_e = 1000$)	0.243	0.101	1.0×10^3
EnKF ($N_e = 3000$)	0.161	0.137	3.0×10^3
EnKF ($N_e = 6000$)	0.127	0.148	6.0×10^3
EnKF (large $N_e = 18500$)	0.111	0.154	1.85×10^4

**Fig. 14** Two-phase model (large number of wells). Left: True log-permeability [$\log m^2$]. Middle: Well configuration. Right: Gelman-rubin diagnostic

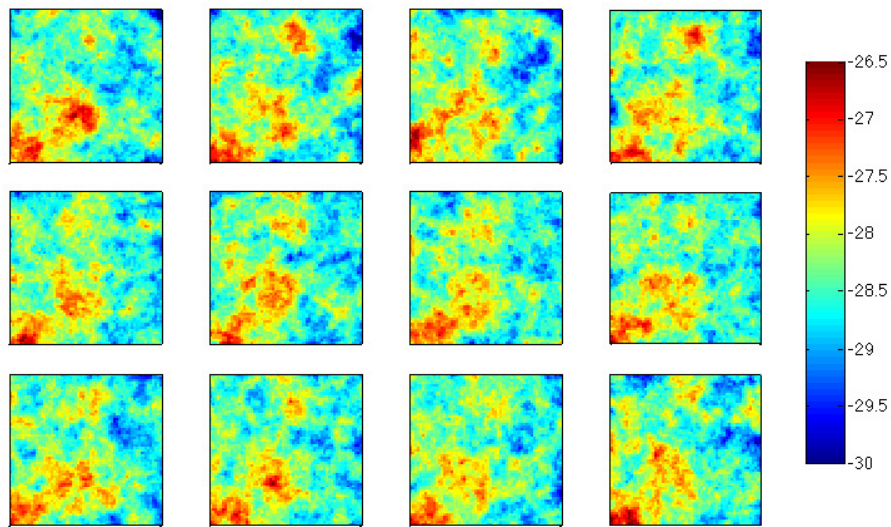


Fig. 15 Two-phase model (small number of wells). Samples from the posterior distribution (characterized with MCMC) [$\log m^2$]

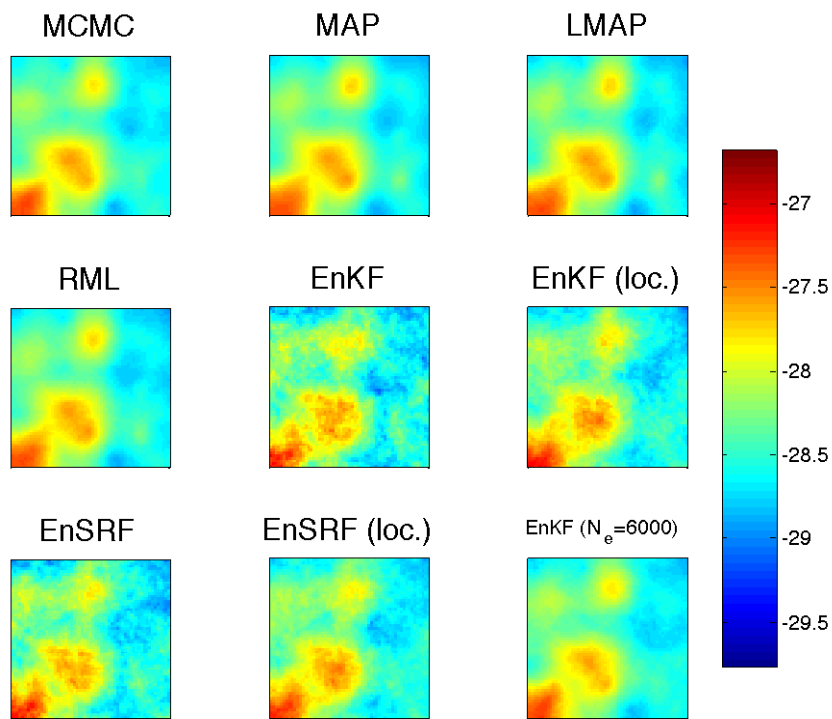


Fig. 16 Two-phase model (large number of wells). Mean of the posterior distribution (characterized with MCMC) and Gaussian approximations [$\log m^2$]

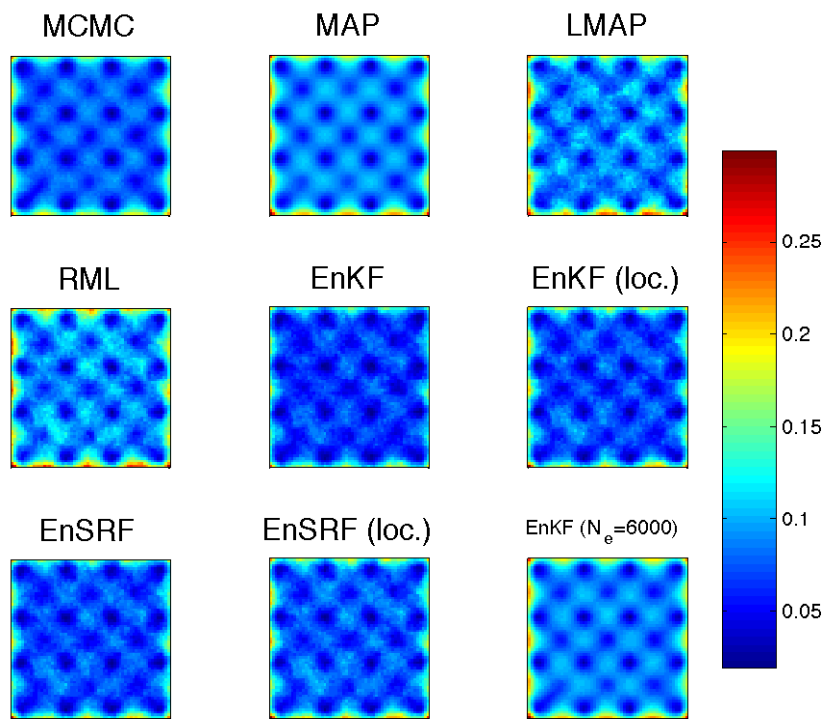


Fig. 17 Two-phase model (large number of wells). Variance of the posterior distribution (characterized with MCMC) and Gaussian approximations [$(\log m^2)^2$]

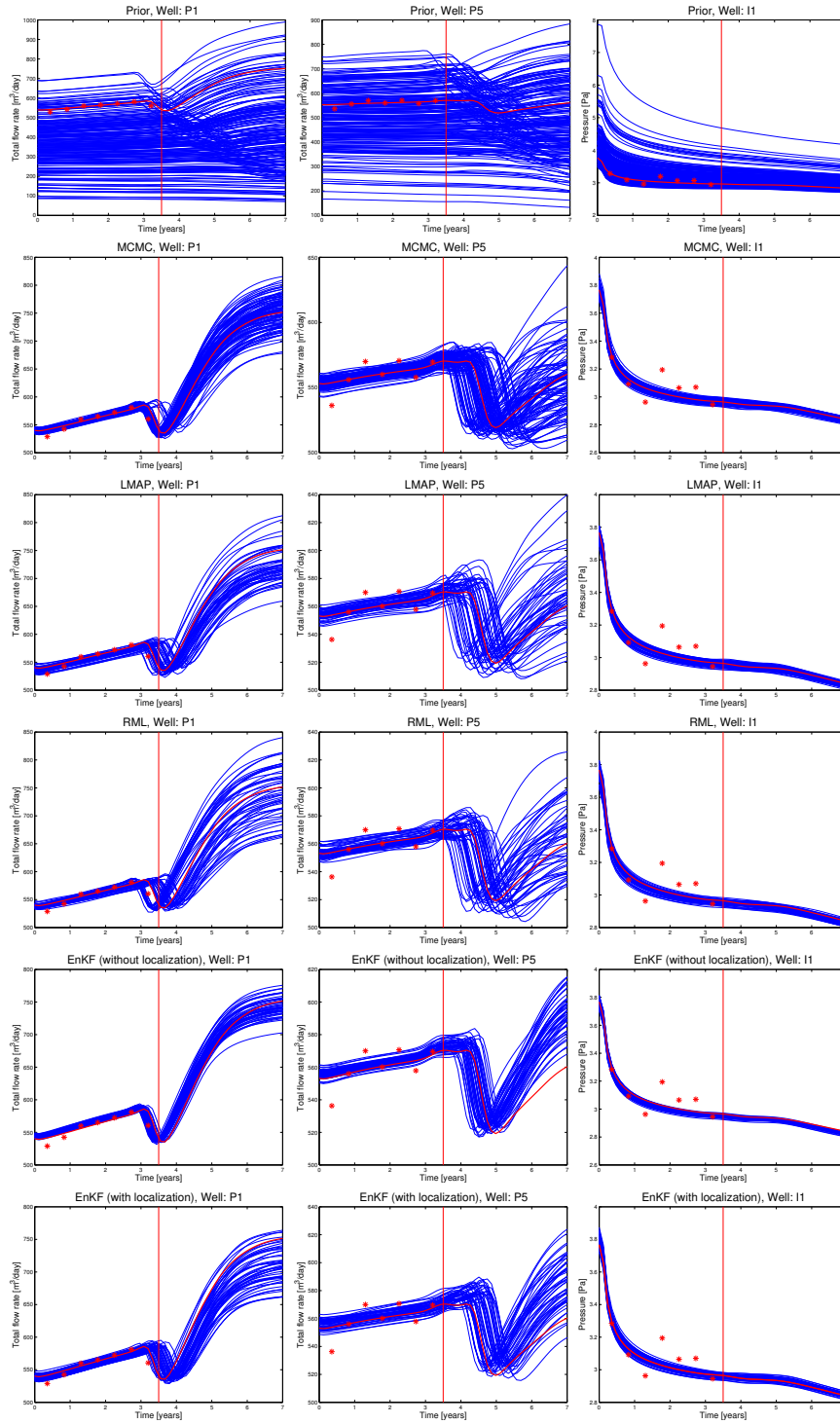


Fig. 18 Two-phase model (large number of wells). Total flow rates from P_1 (left column), P_5 (middle column) and bottom-hole pressure from I_1 (right column) simulated with permeabilities sampled from (top to bottom rows) the prior, the posterior, LMAP, RML, EnKF and EnKF with localization.

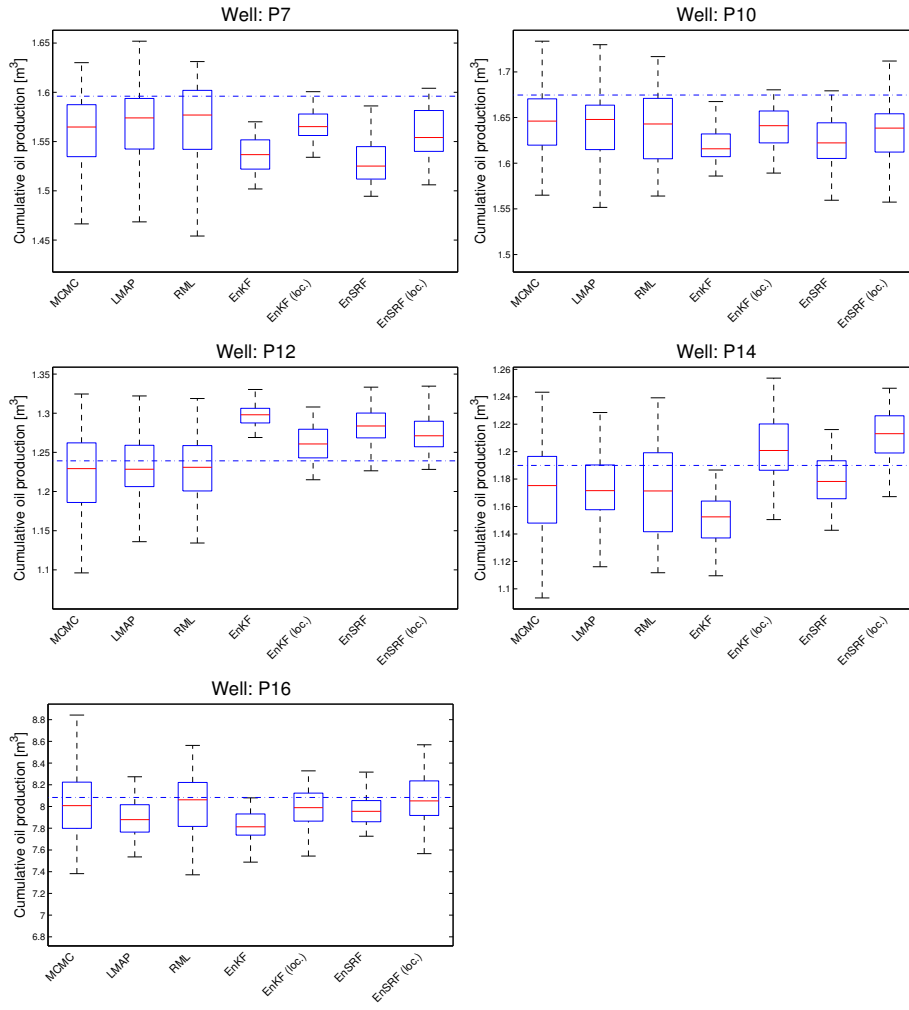


Fig. 19 Two-phase model (large number of wells). Distribution of cumulative oil production at wells $P_7, P_{10}, P_{12}, P_{14}, P_{16}$ at the final time of simulation $t = 10$ years

6 Conclusions

The controlled experiments from the preceding sections enable us to numerically assess the performance of widely used *ad hoc* Gaussian approximations with respect to their ability to correctly reproduce the mean and variance under the true Bayesian posterior distribution. The true posterior is obtained by use of the expensive, but accurate, gold-standard preconditioned Crank-Nicolson Markov Chain Monte Carlo (pCN-MCMC) method. The forward operators associated to the reservoir models under consideration are nonlinear. Therefore, even though a Gaussian prior is considered, the posterior distribution associated to each experiment is non-Gaussian. Indeed, for all our experiments a significant discrepancy between the maximum a posteriori (MAP) estimate and the mean of the posterior distribution was obtained giving clear numerical evidence that the posteriors for the Bayesian data assimilation problems under consideration are not Gaussian. Even in the single-phase reservoir whose associated forward operator is “less” nonlinear (the PDE (1) is linear with respect to p), we find that the relative error in the MAP estimator with respect to the mean of the posterior was 3%. For the two-phase reservoirs, this relative error was 27% and 13% in the case of small and large number of wells, respectively. Thus the problems that we study demonstrate a range of deviations from Gaussianity in the posterior. This makes them a suitable range of test problems for the *ad hoc* algorithms, all of which can be systematically derived in the linear Gaussian scenario, but whose accuracy in the non-Gaussian case is, in general, unclear.

We clearly observe substantial differences in the approximation properties of the posterior distribution with respect to the choice of method, reservoir model and well configuration. For all the experiments conducted here we conclude that, among all the Gaussian approximations under analysis, the linearization around the MAP (LMAP) is arguably the best technique at reproducing the posterior distribution in terms of combined variance and mean. It is interesting to speculate why this might be so, and what it tells us about the posterior. The LMAP algorithm is the only Gaussian approximation which samples from $N(u_{MAP}, C_{MAP})$ where u_{MAP} is the MAP estimate and C_{MAP} the associated covariance matrix defined by (24) and (25), respectively. This suggests that, out of all the Gaussian approximations considered, the posterior distribution in all our examples can be best approximated, in terms of mean and variance, by $N(u_{MAP}, C_{MAP})$. We emphasize that this does not imply that the posterior distribution is Gaussian. Indeed, although this may be the best approximation, errors may still be large.

We recall that all the techniques described in Section 4 produce samples of the posterior distribution in the linear-Gaussian case. In other words, they sample from the exact posterior distribution $N(u_{MAP}, C_{MAP})$. In our experiments, however, we observe clear differences in the approximations obtained with each of the techniques under consideration. For example, note from all our experiments that the randomized maximum likelihood (RML) provided the best approximation of the posterior in terms of the mean. In addition, in the case of single-phase, RML provides a reasonable approximation of the posterior variance (like the one obtained with LMAP). It is worth mentioning that favorable RML results for single-phase reservoirs are also reported in [17]. In contrast, for the two-phase model we find examples where the

error of the RML variance is the largest compared to other Gaussian approximations. These observations are likely to be related to the higher nonlinearity in the two-phase forward operator (17) that results from the nonlinear PDE system (7)-(8). Due to the aforementioned higher nonlinearity, large changes in the absolute values of the log permeability field may not necessarily correspond to large changes in the production data. In fact, production data may typically have smaller sensitivity to the permeability values far from (or in-between) the well locations. On the other hand, we recall from the RML algorithm, that each ensemble member $u_{RML}^{(j)}$ (see equation (27)) produces a model output $G(u_{RML}^{(j)})$ that is close to the perturbed data $y^{(j)}$ while keeping $u_{RML}^{(j)}$ close to the corresponding sample from the prior $u^{(j)}$. Due to the aforementioned small sensitivity of production data to the log-permeability in some regions of the domain, it may be possible that the penalty term $\|u_{RML}^{(j)} - u^{(j)}\|$ in (27) may not provide sufficient constraint to avoid possible large values of $|u_{RML}^{(j)}|$ in the aforementioned regions for which the (perturbed) production data is minimally affected by large value of permeability. Although in our controlled experiment LMAP outperformed the RML in terms of combined mean and variance, it has been reported that RML has the advantage of approximating multimodal distribution for which $N(u_{MAP}, C_{MAP})$ and therefore LMAP is suboptimal. The assessment of techniques where multimodal posterior distribution arises deserves further investigation; however it has not formed part of our studies which have been confined to problems with unimodal posteriors. Moreover, we recall that the computational cost of RML can be amortized if each ensemble member is computed in parallel. Therefore, the cost of the parallel implementation of RML equals the cost of LMAP. Note also that, for very large problems, the factorization of C_{MAP} used in (26) may be computationally prohibitive while the covariance of RML is computed directly from the ensemble at a negligible cost.

For each of our experiments, very poor approximations of the posterior distribution are obtained with the ensemble Kalman filter (EnKF) with a small ensemble size. However, covariance localization leads to a significant reduction in the relative error of the mean and variance with respect to the posterior. Note for example that, in the two-phase model with small number of wells, localization reduces the relative error in the mean and the variance by a factor of two. Additionally, the ensemble Kalman smoother (EnSRF) provides better approximations of the posterior (in terms of mean and variance) than the ones obtained with EnKF. Furthermore, in all our experiments, the ensemble generated with EnSRF with localization provides the best approximation of the posterior in terms of variance. The advantage of using covariance localization as well as using EnSRF instead of EnKF has been widely investigated in terms of reconstructing the truth and/or recovering the truth within the confidence intervals provided by the ensemble approximations. Our results offer now numerical evidence of the advantage of using covariance localization and square root filters for reconstructing the posterior distribution. The choice of covariance localization that provides optimal approximation of the posterior distribution must be further investigated.

Reducing the detrimental effect of sampling error due to the small ensemble size and the possible large amount data is essential in practical applications where a small

number of ensemble members is required to avoid high computational cost in data assimilation. Nonetheless, our results indicate that even for a large ensemble size where presumably sampling error issues are attenuated, we find that EnKF does not converge to the posterior distribution. In fact, as the ensemble sizes increased, the converged Gaussian approximation provided by EnKF resulted in errors of at least 10% both in mean and variance. In addition, the approximations provided with those large size ensembles do not coincide with the approximations provided by either LMAP and RML.

In summary, our study sheds light on various aspects of the *ad hoc* Gaussian approximate filters used in practice to approximate high dimensional posterior distributions on geological reservoir properties. The study has been made possible by use of a fully resolved gold-standard MCMC computation which allows for a clear and well-founded evaluation of the *ad hoc* algorithms. In our opinion more evaluations of this kind will be beneficial in guiding the future evolution of the *ad hoc* Gaussian approximate filters that are so widely used in practice.

Acknowledgements MI, KJHL and AMS gratefully acknowledge the support of EPSRC, ERC, ESA and ONR for various aspects of this work.

References

1. S.I. Aanonsen, G. Naevdal, D.S. Oliver, A.C. Reynolds, , and B. Valles. The Ensemble Kalman Filter in Reservoir Engineering—a Review. *SPE J.*, 14(3):393–412, 2009.
2. B. Arpat, Caers J., and S. Strebelle. Feature-based geostatistics: An application to a submarine channel reservoir. in *Proceedings of the SPE Annual Technical Conference and Exhibition*, San Antonio, Texas, USA, 21-22 February, SPE 77426, 2002.
3. J.W. Barker, M. Cuypers, and L. Holden. Quantifying Uncertainty in Production Forecasts: Another Look at the PUNQ-S3 Problem. *SPE J.*, 6:433–441, December 2001.
4. S.P. Brooks and A. Gelman. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, December 1998.
5. Y. Chen and D. Oliver. Cross-covariances and localization for EnKF in multiphase flow data assimilation. *Computational Geosciences*, 14:579–601, 2010. 10.1007/s10596-009-9174-6.
6. Z. Chen, G. Huan, and Y. MA. *Computational Methods for Multiphase Flows in Porous Media*. Society for Industrial and Applied Mathematics, Philadelphia,PA,U.S.A, 2006.
7. S.L. Cotter, G.O. Roberts, A.M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Submitted*, page arXiv:1202.0709, 2012.
8. C. V. Deutsch. *Geostatistical Reservoir Modeling*. Oxford University Press, Oxford, 2002.
9. Y. Efendiev, A. Datta-Gupta, X. Ma, and B. Mallic. Efficient sampling techniques for uncertainty quantification in history matching using nonlinear error models and ensemble level upscaling techniques. *Water Resources Research*, 45(W11414):11pp, 2009.
10. A. Emerick and A. Reynolds. EnKF-MCMC. in *Proceedings of the SPE EUROPEC/EAGE Annual Conference and Exhibition, Barcelona, Spain, 4-17 June*, SPE 131375, 2010.
11. A. Emerick and A. Reynolds. Combining sensitivities and prior information for covariance localization in the ensemble Kalman filter for petroleum reservoir applications. *Computational Geosciences*, 15:251–269, 2011. 10.1007/s10596-010-9198-y.
12. A. Emerick and A. Reynolds. Combining the Ensemble Kalman Filter with Markov Chain Monte Carlo for improved history matching and uncertainty characterization. in *Proceedings of the SPE Reservoir Simulation Symposium, The Woodlands, Texas, USA, 21-22 February*, SPE 141336, 2012.
13. G. Gao, M. Zafari, and A. Reynolds. Quantifying Uncertainty for the PUNQ-S3 Problem in a Bayesian Setting With RML and EnKF. *SPE J.*, 11:506–515, December 2006.
14. G. Gaspari and S. E. Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554):723–757, 1999.

15. K.J.H. Law and A.M. Stuart. Evaluating data assimilation algorithms. *Monthly Weather Review*, (140):3757–3782, 2012.
16. G. Li and A. Reynolds. Iterative Ensemble Kalman Filters for Data Assimilation. in *Proceedings of the SPE Annual Technical Conference and Exhibition, Anaheim, California, USA, 11-14 Noviembre*, SPE 109808, 2007.
17. N. Liu and D.S Oliver. Evaluation of Monte Carlo methods for assessing uncertainty. *SPE J.*, 8:188–195, 2003.
18. K. Law M. Iglesias and A.M. Stuart. Ensemble Kalman methods for inverse problems. *Submitted*, page arXiv:1202.0709, 2012.
19. X. Ma, A. Al-Harbi, A. Datta-Gupta, and Efendiev Y. An efficient two-stage sampling method for uncertainty quantification in history matching geological models. *SPE J.*, 13(1):7787, 2008.
20. A. C. Reynolds Oliver, D. S. and N. Liu. *Inverse Theory for Petroleum Reservoir Characterization and History Matching*. Cambridge University Press, ISBN: 9780521881517, 1st edition, 2008.
21. D. Oliver and Y. Chen. Recent progress on reservoir history matching: a review. *Computational Geosciences*, 15:185–221, 2011. 10.1007/s10596-010-9194-2.
22. D.S. Oliver, L.B. Cunha, and A.C. Reynolds. Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. *Mathematical Geology*, 29:61–91, 1997.
23. P. Sakov and P.R. Oke. Implications of the form of the ensemble transformation in the ensemble square root filters. *Mon. Wea. Rev.*, 136:10421053, 2008.
24. A.M. Stuart. Inverse problems: a Bayesian perspective. In *Acta Numerica*, volume 19. 2010.
25. R. Tavakoli and A. Reynolds. Monte Carlo simulation of permeability fields and reservoir performance predictions with SVD parameterization in RML compared with EnKF. *Computational Geosciences*, 15:99–116, 2011. 10.1007/s10596-010-9200-8.